Let $\mathbf{x}_1$ and $\mathbf{x}_2$ correspond to distinct eigenvalues λ_1 and λ_2, and let constants k_1 and k_2 be such that

$$k_1\mathbf{x}_1 + k_2\mathbf{x}_2 = \mathbf{0}.$$

Then

$$\mathbf{A}(k_1\mathbf{x}_1 + k_2\mathbf{x}_2) = \mathbf{0},$$

but $\mathbf{A}\mathbf{x}_i = \lambda_i\mathbf{x}_i$, so this is equivalent to

$$k_1\lambda_1\mathbf{x}_1 + k_2\lambda_2\mathbf{x}_2 = \mathbf{0}.$$

Subtracting λ_2 times the first equation from the last result gives

$$(\lambda_1 - \lambda_2)k_1\mathbf{x}_1 = \mathbf{0}.$$

By hypothesis, $\lambda_1 \neq \lambda_2$, so as $\mathbf{x}_1 \neq \mathbf{0}$ it follows that $k_1 = 0$. Using this result in $k_1\mathbf{x}_1 + k_2\mathbf{x}_2 = \mathbf{0}$ shows that $k_2 = 0$, so we have established the linear independence of $\mathbf{x}_1$ and $\mathbf{x}_2$.

To proceed with an inductive proof we now assume that linear independence has been proved for the first $r - 1$ vectors, and show that the rth vector must also be linearly independent. To accomplish this we consider the equation

$$k_1\mathbf{x}_1 + k_2\mathbf{x}_2 + \cdots + k_r\mathbf{x}_r = \mathbf{0}.$$

Premultiplying this equation by $\mathbf{A}$ and reasoning as before, we arrive at the result

$$k_1\lambda_1\mathbf{x}_1 + k_2\lambda_2\mathbf{x}_2 + \cdots + k_r\lambda_r\mathbf{x}_r = \mathbf{0}.$$

Subtracting λ_r times the first equation from the last one gives

$$(\lambda_1 - \lambda_r)k_1\mathbf{x}_1 + (\lambda_2 - \lambda_r)k_2\mathbf{x}_2 + \cdots + (\lambda_{r-1} - \lambda_r)k_{r-1}\mathbf{x}_{r-1} = \mathbf{0}.$$

By the inductive hypothesis $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{r-1}$ are linearly independent, so as $\mathbf{x}_r \neq \mathbf{0}$,

$$(\lambda_1 - \lambda_r)k_1 = (\lambda_2 - \lambda_r)k_2 = \cdots = (\lambda_{r-1} - \lambda_r)k_{r-1} = 0.$$

The eigenvalues are distinct, so the last result can only be true if $k_1 = k_2 = \cdots = k_{r-1} = 0$. Thus $k_r = 0$, and so the vector $\mathbf{x}_r$ is linearly independent of the vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{r-1}$. It has been shown that $\mathbf{x}_1$ and $\mathbf{x}_2$ are linearly independent, so by induction we conclude that the set of vectors $\mathbf{x}_i$ is linearly independent for $i = 1, 2, \ldots, m$.

A matrix $\mathbf{A}$ can have no more than n linearly independent eigenvectors, so when $m = n$ the set of eigenvectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ spans the n-dimensional vector space associated with matrix $\mathbf{A}$ and forms a basis for this space. The proof is complete. ■

algebraic and geometric multiplicity

It can happen that an eigenvalue with **algebraic multiplicity** $r > 1$ only has s different eigenvectors associated with it, where $s < r$, and when this occurs the number s is called the **geometric multiplicity** of the eigenvalue. The set of all eigenvectors associated with an eigenvalue with geometric multiplicity s together with the null vector $\mathbf{0}$ forms what is called the **eigenspace** associated with the eigenvalue. When one or more eigenvalues has a geometric multiplicity that is less than its algebraic multiplicity, it follows directly that the vector space associated with $\mathbf{A}$ must have dimension less than n.

EXAMPLE 4.1 Find the characteristic polynomial, the eigenvalues, and the eigenvectors of the matrix

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & -1 \\ 3 & 2 & -3 \\ 3 & 1 & -2 \end{bmatrix}.$$

Solution The characteristic polynomial $P_3(\lambda)$ is given by

$$P_3(\lambda) = \begin{vmatrix} 2-\lambda & 1 & -1 \\ 3 & 2-\lambda & -3 \\ 3 & 1 & -2-\lambda \end{vmatrix},$$

and after expanding the determinant we find that

$$P_3(\lambda) = -\lambda^3 + 2\lambda^2 + \lambda - 2.$$

The characteristic equation $P_3(\lambda) = 0$ is

$$\lambda^3 - 2\lambda^2 - \lambda + 2 = 0,$$

and inspection shows it has the roots 2, 1, and -1. So the *eigenvalues* of $\mathbf{A}$ are $\lambda_1 = 2, \lambda_2 = 1$, and $\lambda_3 = -1$, and as these roots are all distinct (there are no repeated roots), each has an algebraic and geometric multiplicity of 1 (each is a single root). The set of numbers $-1, 1, 2$ forms the *spectrum* of matrix $\mathbf{A}$. As the *spectral radius* R of a matrix is defined as the largest of the moduli of the eigenvalues, we see that $R = 2$.

To find the eigenvectors $\mathbf{x}_i$ of $\mathbf{A}$ corresponding to the eigenvalues $\lambda = \lambda_i$, for $i = 1, 2, 3$, it will be necessary to solve the homogeneous system of algebraic equations

$$(\mathbf{A} - \lambda_i \mathbf{I})\mathbf{x}_i = \mathbf{0} \quad \text{for } i = 1, 2, 3,$$

where $\mathbf{x}_i = [x_1, x_2, x_3]^{\mathrm{T}}$.

Case $\lambda_1 = 2$

The system of equations to be solved is

$$\begin{bmatrix} 2-2 & 1 & -1 \\ 3 & 2-2 & -3 \\ 3 & 1 & -2-2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix},$$

and this matrix equation is equivalent to the set of three linear algebraic equations

$$x_2 - x_3 = 0, \quad 3x_1 - 3x_3 = 0, \quad \text{and} \quad 3x_1 + x_2 - 4x_3 = 0.$$

The first two equations are equivalent, so only one of the first two equations and the third equation are linearly independent. Solving the last two equations for x_1 and x_2 in terms of x_3, we find that $x_1 = x_2 = x_3$, so setting $x_3 = k_1$ where k_1 is an arbitrary real number (a parameter) shows that the eigenvector $\mathbf{x}_1$ corresponding to the eigenvalue $\lambda_1 = 2$ is given by

$$\mathbf{x}_1 = \begin{bmatrix} k_1 \\ k_1 \\ k_1 \end{bmatrix} = k_1 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

As k_1 is an arbitrary parameter, for convenience we set $k_1 = 1$ and as a result obtain the eigenvector

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

Case $\lambda_2 = 1$

This time the system of equations to be solved to find the eigenvector $\mathbf{x}_2$ is

$$\begin{bmatrix} 2-1 & 1 & -1 \\ 3 & 2-1 & -3 \\ 3 & 1 & -2-1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix},$$

and this is equivalent to the three linear algebraic equations

$$x_1 + x_2 - x_3 = 0, \quad 3x_1 + x_2 - 3x_3 = 0, \quad \text{and} \quad 3x_1 + x_2 - 3x_3 = 0.$$

The last two equations are identical, so we must solve for x_1, x_2, and x_3 using the first two equations. It is easily seen from these two equations that $x_2 = 0$ and $x_1 = x_3$, so setting $x_1 = k_2$, where k_2 is an arbitrary real number (a parameter), gives

$$\mathbf{x}_2 = k_2 \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}.$$

Making the arbitrary choice $k_2 = 1$ shows that the eigenvector $\mathbf{x}_2$ corresponding to $\lambda_2 = 1$ is

$$\mathbf{x}_2 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}.$$

Case $\lambda_3 = -1$

Setting $\lambda = \lambda_3$, and proceeding as before, shows that the elements of the eigenvector $\mathbf{x}_3$ must satisfy the three equations

$$3x_1 + x_2 - x_3 = 0, \quad 3x_1 + 3x_2 - 3x_3 = 0, \quad \text{and} \quad 3x_1 + x_2 - x_3 = 0,$$

with the solution $x_1 = 0, x_2 = x_3 = k_3$, where k_3 is an arbitrary real number (a parameter). Making the arbitrary choice $k_3 = 1$ allows the eigenvector $\mathbf{x}_3$ to be written as

$$\mathbf{x}_3 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}.$$

We have shown that matrix $\mathbf{A}$ has the three distinct eigenvalues $\lambda_1 = 2, \lambda_2 = 1$, and $\lambda_3 = -1$, corresponding to which there are the three eigenvectors

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \quad \text{and} \quad \mathbf{x}_3 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}.$$

These three eigenvectors form a basis for the three-dimensional vector space associated with $\mathbf{A}$. ■

As the eigenvectors $\mathbf{x}$ of matrix $\mathbf{A}$ satisfy the homogeneous equation (2), they can be multiplied by an arbitrary nonzero number K, which is either positive or negative, and still remain an eigenvector. This property is used to *scale* the eigenvectors of $\mathbf{A}$ to produce what are called **normalized** eigenvectors. This scaling is used in numerical calculations involving the iteration of eigenvectors, because without normalization the elements of $\mathbf{x}$ may either grow or diminish in absolute value after each stage of the calculation, leading to a progressive loss of accuracy.

Normalization of eigenvectors

a frequently used way of normalizing eigenvectors

Various normalizations are in use. The most common one for eigenvectors with real elements involves scaling the eigenvector so that the square root of the sum of the squares of its elements is 1. So, for example, if

$$\mathbf{x} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}, \quad \text{the normalizing factor} \quad K = \frac{1}{(a^2 + b^2 + c^2)^{1/2}} \tag{6}$$

and the normalized eigenvector $\hat{\mathbf{x}}$ becomes

$$\hat{\mathbf{x}} = \begin{bmatrix} a/(a^2 + b^2 + c^2)^{1/2} \\ b/(a^2 + b^2 + c^2)^{1/2} \\ c/(a^2 + b^2 + c^2)^{1/2} \end{bmatrix}. \tag{7}$$

When the eigenvectors in Example 4.1 are normalized in this way, they become

$$\hat{\mathbf{x}}_1 = \begin{bmatrix} 1/\sqrt{3} \\ 1/\sqrt{3} \\ 1/\sqrt{3} \end{bmatrix}, \quad \hat{\mathbf{x}}_2 = \begin{bmatrix} 1/\sqrt{2} \\ 0 \\ 1/\sqrt{2} \end{bmatrix}, \quad \text{and} \quad \hat{\mathbf{x}}_3 = \begin{bmatrix} 0 \\ 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}.$$

EXAMPLE 4.2

Find the characteristic polynomial, eigenvalues, and eigenvectors of the matrix

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 1 & 1 \\ -1 & 2 & 0 & 1 \\ -1 & 0 & 2 & 1 \\ 1 & 0 & -1 & 0 \end{bmatrix}.$$

Solution The determinant defining the characteristic polynomial is

$$P_4(\lambda) = \begin{bmatrix} -\lambda & 0 & 1 & 1 \\ -1 & 2-\lambda & 0 & 1 \\ -1 & 0 & 2-\lambda & 1 \\ 1 & 0 & -1 & -\lambda \end{bmatrix},$$

and after the determinant is expanded the characteristic equation $P_4(\lambda) = 0$ is found to be

$$P_4(\lambda) = \lambda(\lambda^3 - 4\lambda^2 + 5\lambda - 2) = 0.$$

Clearly, $\lambda = 0$ is a root of $P_4(\lambda) = 0$, and inspection shows the other three roots to be 1, 1, and 2. So the eigenvalues of $\mathbf{A}$ are $\lambda_1 = 0$, $\lambda_2 = 1$, $\lambda_3 = 1$, and $\lambda_4 = 2$. In this

case $\lambda_2 = \lambda_3 = 1$, so the eigenvalue 1 has algebraic multiplicity 2, and the remaining two eigenvalues each have an algebraic multiplicity of 1. To find the eigenvectors corresponding to these eigenvalues we proceed as in Example 4.1.

Case $\lambda_1 = 0$

Setting $\lambda = \lambda_1 = 0$ in $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = 0$ leads to the four equations

$$x_3 + x_4 = 0, \quad -x_1 + 2x_2 + x_4 = 0, \quad -x_1 + 2x_3 + x_4 = 0, \quad \text{and} \quad x_1 - x_3 = 0.$$

Proceeding as before we find that $x_1 = x_2 = x_3 = -x_4$, so solving for x_1, x_2, and x_3 in terms of x_4, and setting $x_4 = 1$ (an arbitrary choice), shows the eigenvector $\mathbf{x}_1$ to be

$$\mathbf{x}_1 = \begin{bmatrix} -1 \\ -1 \\ -1 \\ 1 \end{bmatrix}.$$

Case $\lambda_2 = \lambda_3 = 1$

The eigenvalue 1 has algebraic multiplicity 2, so we must attempt to find two *different* eigenvectors that correspond to the single eigenvalue $\lambda = 1$. Setting $\lambda = 1$ in $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$ leads to the four equations

$$-x_1 + x_3 + x_4 = 0, \quad -x_1 + x_2 + x_4 = 0, \quad -x_1 + x_3 + x_4 = 0, \quad x_1 - x_3 - x_4 = 0.$$

The first, third, and fourth equations are identical, so x_1, x_2, x_3, and x_4 must be determined from the two equations

$$-x_1 + x_3 + x_4 = 0 \quad \text{and} \quad -x_1 + x_2 + x_4 = 0.$$

As there are four unknown quantities x_1, x_2, x_3, and x_4, and only two equations relating them, it will only be possible to solve for two of these quantities in terms of the remaining two. The equations show that $x_2 = x_3$ and $x_4 = x_1 - x_3$, so choosing to solve for x_3 and x_4 in terms of x_1 and x_2 by setting $x_1 = \alpha$ and $x_2 = \beta$, with α and β arbitrary constants, shows that the eigenvectors $\mathbf{x}_2$ and $\mathbf{x}_3$ are both of the form

$$\mathbf{x}_{2,3} = \begin{bmatrix} \alpha \\ \beta \\ \beta \\ \alpha - \beta \end{bmatrix}.$$

It is possible to obtain two *different* eigenvectors from this last result by choosing two different pairs of values for the arbitrary parameters α and β. We will define $\mathbf{x}_2$ by setting $\alpha = 1$ and $\beta = 1$, and $\mathbf{x}_3$ by setting $\alpha = 1$ and $\beta = 0$, and as a result we find that

$$\mathbf{x}_2 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \end{bmatrix} \quad \text{and} \quad \mathbf{x}_3 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}.$$

Had other choices of the parameters α and β been made, two different eigenvectors would have been produced.

Case $\lambda_4 = 2$

Setting $\lambda = \lambda_4 = 2$ in $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = 0$ leads to the four equations

$$-2x_1 + x_3 + x_4 = 0, \quad -x_1 + x_4 = 0, \quad -x_1 + x_4 = 0, \quad x_1 - x_3 - 2x_4 = 0.$$

These equations have the solution $x_1 = x_3 = x_4 = 0$, with no condition being imposed on x_2. For simplicity we choose to set $x_2 = 1$ to obtain

$$\mathbf{x}_4 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}.$$

In this example, the eigenvalue 1 has algebraic multiplicity 2, and two different eigenvectors can be associated with it, so the geometric multiplicity of the eigenvalue is also 2. The four eigenvectors $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$, and $\mathbf{x}_4$ form a basis for the four-dimensional vector space associated with matrix $\mathbf{A}$.

Had different values been used for α and β, the basis vectors for this vector space would have been different, though the vector space itself would have remained the same because linear combinations of basis vectors will produce an equivalent set of basis vectors.

The *spectrum* of $\mathbf{A}$ is the set of numbers 0, 1, 2, and the *spectral radius* of $\mathbf{A}$ is seen to be $R = 2$. ■

EXAMPLE 4.3

Show that the matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

has three eigenvalues, but only two linearly independent eigenvectors.

Solution The characteristic polynomial

$$P_3(\lambda) = \begin{vmatrix} 1-\lambda & 1 & 0 \\ 0 & 1-\lambda & 0 \\ 0 & 0 & -\lambda \end{vmatrix},$$

and after expanding the determinant the characteristic equation $P_3(\lambda) = 0$ becomes

$$P_3(\lambda) = -\lambda(1-\lambda)^2 = 0.$$

The eigenvalue $\lambda_1 = 0$ occurs with algebraic multiplicity 1 and the eigenvalue $\lambda_2 = \lambda_3 = 1$ occurs with algebraic multiplicity 2.

The equations determining the eigenvector $\mathbf{x}_1$, corresponding to the eigenvalue $\lambda = \lambda_1 = 0$, are

$$x_1 + x_2 = 0 \quad \text{and} \quad x_2 = 0,$$

so $x_1 = x_2 = 0$ and x_3 is arbitrary. Setting $x_3 = 1$ gives

$$\mathbf{x}_1 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

The equations determining $\mathbf{x}_2$ and $\mathbf{x}_3$, corresponding to $\lambda = \lambda_2 = \lambda_3 = 1$, are

$$x_1 = k(\text{arbitrary}) \quad \text{and} \quad x_2 = x_3 = 0,$$

so setting $k = 1$, we find that the eigenvalue $\lambda_2 = \lambda_3 = 1$ with algebraic multiplicity 2 only has associated with it the *single* eigenvector

$$\mathbf{x}_{2,3} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

So the algebraic multiplicity of the eigenvalue $\lambda = 1$ is 2, but its geometric multiplicity is 1. The *spectrum* of $\mathbf{A}$ is the set of numbers 0, 1, so the *spectral radius* of $\mathbf{A}$ is $R = 1$. ■

The eigenvalues of a diagonal matrix can be found immediately, and the corresponding eigenvectors take on a particularly simple form. Let $\mathbf{D}$ be the $n \times n$ diagonal matrix

$$\mathbf{D} = \begin{bmatrix} a_1 & 0 & 0 & \cdots & 0 \\ 0 & a_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & a_n \end{bmatrix},$$

with entries $a_1, a_2, \ldots, a_n$ on its leading diagonal, not all of which are zero, and zeros elsewhere. Then it is easily seen that the eigenvalues of $\mathbf{D}$ are $\lambda_1 = a_1, \lambda_2 = a_2, \ldots, \lambda_n = a_n$. The eigenvector $\mathbf{x}_i$ corresponding to the eigenvalue $\lambda_i = a_i$ becomes an n-element column vector in which only the ith element is nonzero. It is not difficult to show that this result remains true whatever the algebraic multiplicity of an eigenvalue, so *every* diagonal $n \times n$ matrix has n eigenvectors of this form. For convenience, the ith element in $\mathbf{x}_i$ is usually taken to be 1 so, for example, the matrix

$$\mathbf{A} = \begin{bmatrix} 3 & 0 & 0 \\ 0 & -5 & 0 \\ 0 & 0 & 4 \end{bmatrix}$$

has eigenvalues $\lambda_1 = 3$, $\lambda_2 = -5$, and $\lambda_3 = 4$ and eigenvectors

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \text{and} \quad \mathbf{x}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

Similarly, the diagonal matrix

$$\mathbf{A} = \begin{bmatrix} -2 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix}$$

has an eigenvalue $\lambda_1 = -2$ with multiplicity 1 and a double eigenvalue $\lambda_2 = \lambda_3 = 4$ with multiplicity 2, but the matrix still has the three distinct eigenvectors

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \text{and} \quad \mathbf{x}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

When the degree of the characteristic equation of a matrix exceeds 2, its roots must usually be found by means of a numerical technique. In such circumstances the next theorem provides a simple and useful check for the values of the eigenvalues that have been computed.

THEOREM 4.2

a check on the sum of the eigenvectors

The sum of eigenvalues Let the $n \times n$ matrix $\mathbf{A}[a_{ij}]$ have the n eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$, which may be either real or complex. Then

$$\lambda_1 + \lambda_2 + \cdots + \lambda_n = (-1)^{n-1}(a_{11} + a_{22} + \cdots + a_{nn}) = (-1)^{n-1}\text{tr}(\mathbf{A}).$$

Proof As the multiplication of a column of a matrix by a number k is equivalent to multiplication of its determinant by k, we can write

$$P_n(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I}) = (-1)^n \det(\lambda\mathbf{I} - \mathbf{A}).$$

Expanding the determinant on the right in terms of the elements of the first column and separating out the factors that can give rise to the terms in λ^n and λ^{n-1}, we arrive at the result

$$P_n(\lambda) = (-1)^n\{(\lambda - a_{11})(\lambda - a_{22}) \cdots (\lambda - a_{nn}) + Q_{n-2}(\lambda)\},$$

where $Q_{n-2}(\lambda)$ is a polynomial in λ of degree $n - 2$.

Identifying the coefficients of λ^n and λ^{n-1} in the expression for $P_n(\lambda)$ shows that

$$P_n(\lambda) = (-1)^n\{\lambda^n - (a_{11} + a_{22} + \cdots + a_{nn})\lambda^{n-1} + \cdots + \text{constant} + Q_{n-2}(\lambda)\}.$$

An equivalent expression for $P_n(\lambda)$ can be obtained by expanding it in terms of its factors $(\lambda - \lambda_1), (\lambda - \lambda_2), \ldots, (\lambda - \lambda_n)$ to obtain

$$\begin{aligned} P_n(\lambda) &= (-1)^n(\lambda - \lambda_1)(\lambda - \lambda_2) \cdots (\lambda - \lambda_n) \\ &= (-1)^n\{\lambda^n - (\lambda_1 + \lambda_2 + \cdots + \lambda_n)\lambda^{n-1} + \cdots + \text{constant}\}. \end{aligned}$$

The statement of the theorem then follows by comparing the coefficients of λ^{n-1} in the two different expressions for $P_n(\lambda)$, where it will be recalled that the **trace** of an $n \times n$ matrix $\mathbf{A}[a_{ij}]$, written $\text{tr}(\mathbf{A})$, is the sum of the elements on its leading diagonal, so that $\text{tr}(\mathbf{A}) = a_{11} + a_{22} + \cdots + a_{nn}$. ■

EXAMPLE 4.4

Use Theorem 4.2 to check the eigenvalues of the matrices in Examples 4.1 and 4.2.

Solution In Example 4.1, $\lambda_1 = 2$, $\lambda_2 = 1$, and $\lambda_3 = -1$, so $\lambda_1 + \lambda_2 + \lambda_3 = 2$, and $\text{tr}(\mathbf{A}) = 2 + 2 - 2 = 2$, so the result of Theorem 4.2 is verified. Similarly, in Example 4.2, $\lambda_1 = 0$, $\lambda_2 = 1$, $\lambda_3 = 1$, and $\lambda_4 = 2$, so $\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 4$, and $\text{tr}(\mathbf{A}) = 0 + 2 + 2 + 0 = 4$, showing that the result of Theorem 4.2 is again verified. ■

EXAMPLE 4.5

Find the characteristic polynomial, eigenvalues, and eigenvectors of

$$\mathbf{A} = \begin{bmatrix} -1 - 2i & -1 - i & 2 + 2i \\ -4i & -i & 4i \\ -1 - 3i & -1 - i & 2 + 3i \end{bmatrix},$$

and use Theorem 4.2 to check the eigenvalues.

Solution This matrix has complex elements. Expanding $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$ shows that the characteristic polynomial $P_3(\lambda)$ is

$$P_3(\lambda) = \lambda^3 - \lambda^2 + \lambda - 1.$$

Inspection shows the eigenvalues determined by $P_3(\lambda) = 0$ to be $\lambda_1 = 1$, $\lambda_2 = i$, and $\lambda_3 = -i$. Finding the eigenvectors, as in Example 4.1, gives

$$(\lambda_1 = 1)\,\mathbf{x}_1 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \quad (\lambda_2 = i)\,\mathbf{x}_2 = \begin{bmatrix} 0 \\ 1 \\ 1/2 \end{bmatrix}, \quad \text{and} \quad (\lambda_3 = -i)\,\mathbf{x}_3 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

In this example, although the matrix $\mathbf{A}$ has complex elements, the characteristic polynomial has real coefficients, and one of its zeros (an eigenvalue) is real and its other two zeros (eigenvalues) are complex conjugates. The test in Theorem 4.2 is satisfied because $\text{tr}(\mathbf{A}) = \lambda_1 + \lambda_2 + \lambda_3 = \text{tr}(\mathbf{A}) = 1$. ■

Complex eigenvalues arise in numerous applications of matrices, and when this happens it is often useful to have qualitative information about a region in the complex plane that contains all of the eigenvalues, without the necessity of computing their actual values. This form of approach is particularly useful when the coefficients of a polynomial are not specific, and all that is known is that they lie within given intervals or, if complex, that the modulus of each is bounded by a given number.

Another need for this type of information occurs when working with systems of linear differential equations, because it will be seen in Chapter 6 that the roots of a characteristic polynomial equation determine the form of the general solution of a homogeneous system. Roots of the form $\alpha + i\beta$ will be seen to lead to real solutions of the form $e^{\alpha t}\sin\beta t$ and $e^{\alpha t}\cos\beta t$, and these solutions will only remain bounded (stable) as $t \to +\infty$ if the real part of every root is negative. This means that the qualitative knowledge that all of the roots lie to the left of the imaginary axis will be sufficient to ensure that the solution remains finite (is stable) as $t \to +\infty$.

The theorem that follows is the simplest of many similar results that are available, all of which provide information about regions in the complex plane where all of the zeros of a characteristic polynomial are located. Two other results are to be found in the exercise set at the end of this section; the one called the **Routh–Hurwitz stability criterion** is particularly useful when working with systems of linear differential equations.

Although the theorem to be proved in this section identifies a region less precisely than many similar theorems, it has been included to illustrate how such regions can be found, and also because the derivation of the result is elementary. The proof only uses the basic properties of complex numbers extending as far as the triangle inequality.

THEOREM 4.3

finding a region that contains all the eigenvalues

The Gerschgorin circle theorem Let $\mathbf{A}[a_{ij}]$ be an $n \times n$ matrix, and define the circles $C_1, C_2, \ldots, C_n$ in the complex plane such that circle C_r has its center at a_{rr} and the radius

$$\rho_r = \sum_{j=1,\, j\neq r}^{n} |a_{rj}| = |a_{r1}| + |a_{r2}| + \cdots + |a_{r,r-1}| + |a_{r,r+1}| + \cdots + |a_{rn}|.$$

Then each of the eigenvalues of $\mathbf{A}$ lies in at least one of these circles.

Proof The rth equation of $\mathbf{Ax} = \lambda\mathbf{x}$ is

$$a_{r1}x_1 + \cdots + a_{r,r-1}x_{r-1} + (a_{rr} - \lambda)x_r + a_{r,r+1}x_{r+1} + \cdots + a_{rn}x_n = 0.$$

Solving for $(a_{rr} - \lambda)$, taking the modulus of the result, and making repeated use of the triangle inequality $|a + b| \leq |a| + |b|$, where a and b are arbitrary complex numbers, leads to the inequality

$$|\lambda - a_{rr}| < \sum_{j=1,\ j\neq r}^{n} |a_{rj}||x_j|/|x_r|, \quad \text{for } r = 1, 2, \ldots, n.$$

We now choose x_r to be the element of $\mathbf{x}$ with the largest modulus, so that $|x_j|/|x_r| \leq 1$ for $r = 1, 2, \ldots, n$. The statement of the theorem is obtained from the inequality involving $|\lambda - a_{rr}|$ by replacing each term $|x_j|/|x_r|$ on the right by 1, and then repeating the argument for $r = 1, 2, \ldots, n$. ■

EXAMPLE 4.6 Apply the Gerschgorin circle theorem to Example 4.1.

Solution Circle C_1 has its center at the point $a_{11} = (2, 0)$ and its radius $\rho_1 = |a_{12}| + |a_{13}| = 1 + 1 = 2$. Circle C_2 has its center at the point $a_{22} = (2, 0)$ and its radius $\rho_2 = |a_{21}| + |a_{23}| = 3 + 3 = 6$, while circle C_3 has its center at the point $a_{33} = (-2, 0)$ and its radius $\rho_3 = |a_{31}| + |a_{32}| = 3 + 1 = 4$.

Consequently, the Gerschgorin circle theorem asserts that all the eigenvalues of $\mathbf{A}$ lie in the region of the complex plane enclosed by these three circles. The circles are shown in Fig. 4.1 together with the locations of the three eigenvalues 2, 1, and -1. ■

Physical problems that give rise to matrices with real coefficients often do so in the form of real valued symmetric matrices. These matrices have a number of useful properties that we will examine after first introducing the notions of the *inner product* and *norm* of a matrix vector, and then *orthogonal* and *orthonormal* sets of matrix vectors.

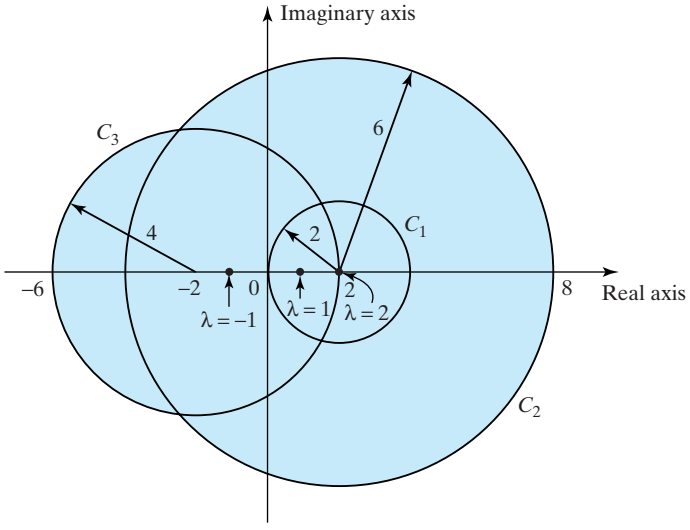


FIGURE 4.1 The Gerschgorin circles for Example 4.1.

inner products, the norm, orthogonal and orthonormal sets of vectors

Inner product of vectors

Let $\mathbf{u}$ and $\mathbf{v}$ be two n-element matrix vectors (row or column) with the respective elements $u_1, u_2, \ldots, u_n$ and $v_1, v_2, \ldots, v_n$. Then their **dot** or **inner product**, denoted here by $\mathbf{u} \cdot \mathbf{v}$ but elsewhere often by $\langle \mathbf{u}, \mathbf{v} \rangle$, is defined as

$$\mathbf{u} \cdot \mathbf{v} = u_1 v_1 + u_2 v_2 + \cdots + u_n v_n. \tag{8}$$

Norm of a vector

The **norm** of an n-element vector $\mathbf{w}$ (row or column) with elements w_1, $w_2, \ldots, w_n$, written $\|\mathbf{w}\|$, is defined as $(\mathbf{w} \cdot \mathbf{w})^{1/2}$, and so is given by

$$\|\mathbf{w}\| = \left(w_1^2 + w_2^2 + \cdots + w_n^2\right)^{1/2}. \tag{9}$$

We now use the matrix norm to introduce the idea of the *orthogonality* of sets of matrix vectors, and then to show how such sets can be replaced by an equivalent *orthonormal* set of vectors.

Orthogonal and orthonormal sets of vectors

Let $\mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u}_n$ be a set of n-element vectors (row or column). Then the set is said to be **orthogonal** if

$$\mathbf{u}_i \cdot \mathbf{u}_j = \begin{cases} 0 & \text{for } i \neq j, \\ \|\mathbf{u}_i\|^2 & \text{for } i = j, \end{cases} \tag{10}$$

and to be **orthonormal** if, in addition to being orthogonal, the norm of each vector is 1, so that $\|\mathbf{u}_i\| = 1$ for $i = 1, 2, \ldots, n$. This means that the set of vectors $\mathbf{u}_1, \mathbf{u}_2 \ldots, \mathbf{u}_n$ will form an *orthonormal* set if

$$\mathbf{u}_i \cdot \mathbf{u}_j = \begin{cases} 0 & \text{for } i \neq j, \\ \|\mathbf{u}_i\|^2 = 1 & \text{for } i = j. \end{cases} \tag{11}$$

EXAMPLE 4.7

Given the sets of vectors

(a)

$$\mathbf{u}_1 = \begin{bmatrix} 1 \\ 2 \\ -2 \end{bmatrix}, \quad \mathbf{u}_2 = \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix} \text{ and } \mathbf{u}_3 = \begin{bmatrix} -2 \\ 2 \\ 1 \end{bmatrix},$$

and

(b)

$$\mathbf{u}_1 = [1/4, \sqrt{3}/4, \sqrt{3}/2], \quad \mathbf{u}_2 = [\sqrt{3}/2, -1/2, 0], \quad \mathbf{u}_3 = [\sqrt{3}/4, 3/4, -1/2],$$

show the vectors in set (a) are orthogonal and convert them to an orthonormal set, and that those in set in (b) are orthonormal.

Solution

(a) $\mathbf{u}_1 \cdot \mathbf{u}_2 = 1.2 + 2.1 - 2.2 = 0$ and, similarly, $\mathbf{u}_1 \cdot \mathbf{u}_3 = \mathbf{u}_2 \cdot \mathbf{u}_3 = 0$, and $\|\mathbf{u}_1\| = \|\mathbf{u}_2\| = \|\mathbf{u}_3\| = \sqrt{9} = 3$. So the set is orthogonal but *not* orthonormal, because the vector norms are not all equal to 1. To convert the set into an orthonormal set, it is only necessary to divide each vector by its norm to arrive at the equivalent *orthonormal* set

$$\hat{\mathbf{u}}_1 = \begin{bmatrix} 1/3 \\ 2/3 \\ -2/3 \end{bmatrix}, \quad \hat{\mathbf{u}}_2 = \begin{bmatrix} 2/3 \\ 1/3 \\ 2/3 \end{bmatrix}, \quad \text{and} \quad \hat{\mathbf{u}}_3 = \begin{bmatrix} -2/3 \\ 2/3 \\ 1/3 \end{bmatrix}.$$

(b) Proceeding as in (a) we have $\mathbf{u}_1 \cdot \mathbf{u}_2 = \mathbf{u}_1 \cdot \mathbf{u}_3 = \mathbf{u}_2 \cdot \mathbf{u}_3 = 0$, showing that the set is orthogonal. However, $\|\mathbf{u}_1\| = \|\mathbf{u}_2\| = \|\mathbf{u}_3\| = 1$, so the set is also orthonormal. ■

THEOREM 4.4

properties of eigenvalues and eigenvectors of symmetric matrices

Eigenvalues and eigenvectors of a symmetric matrix Let $\mathbf{A}$ be an $n \times n$ real symmetric matrix. Then

(i) the eigenvalues of $\mathbf{A}$ are all real;

(ii) the eigenvectors of $\mathbf{A}$ corresponding to distinct eigenvalues are mutually orthogonal.

Proof We start by observing that if $\mathbf{x}$ and $\mathbf{y}$ are two n-element column vectors the product $\mathbf{y}^T\mathbf{A}\mathbf{x}$ is a scalar, and so is equal to its transpose. Thus, $\mathbf{y}^T\mathbf{A}\mathbf{x} = (\mathbf{y}^T\mathbf{A}\mathbf{x})^T = \mathbf{x}^T\mathbf{A}^T\mathbf{y}$, but as $\mathbf{A}$ is symmetric $\mathbf{A}^T = \mathbf{A}$, so that $\mathbf{y}^T\mathbf{A}\mathbf{x} = \mathbf{x}^T\mathbf{A}^T\mathbf{y}$.

To prove (i), let λ be an eigenvalue of $\mathbf{A}$ with the corresponding eigenvector $\mathbf{x}$. Then

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}.$$

Taking the complex conjugate of this result and using the fact that $\mathbf{A}$ is real valued, so that $\overline{\mathbf{A}} = \mathbf{A}$, gives

$$\mathbf{A}\overline{\mathbf{x}} = \overline{\lambda}\overline{\mathbf{x}}.$$

This shows that $\overline{\lambda}$ is an eigenvalue of $\mathbf{A}$ with the associated eigenvector $\overline{\mathbf{x}}$. If we now premultiply this result by $\mathbf{x}^T$, we obtain the scalar equation

$$\mathbf{x}^T\mathbf{A}\overline{\mathbf{x}} = \overline{\lambda}\mathbf{x}^T\overline{\mathbf{x}},$$

but premultiplying the original eigenvalue equation by $\overline{\mathbf{x}}^T$ gives

$$\overline{\mathbf{x}}^T\mathbf{A}\mathbf{x} = \lambda\overline{\mathbf{x}}^T\mathbf{x}.$$

Using the result $\mathbf{x}^T\mathbf{A}\overline{\mathbf{x}} = \overline{\mathbf{x}}^T\mathbf{A}\mathbf{x}$ then shows that $\lambda\overline{\mathbf{x}}^T\mathbf{x} = \overline{\lambda}\mathbf{x}^T\overline{\mathbf{x}}$, but $\overline{\mathbf{x}}^T\mathbf{x} = \mathbf{x}^T\overline{\mathbf{x}}$ so $\lambda = \overline{\lambda}$, which is only possible if λ is real. This has established the first part of the theorem.

To prove (ii) we must show that if $\mathbf{x}_r$ and $\mathbf{x}_s$ are eigenvectors of $\mathbf{A}$ corresponding to the distinct eigenvalues λ_r and λ_s, with $r \neq s$, then $\mathbf{x}_r \cdot \mathbf{x}_s = 0$, which is equivalent to the condition $\mathbf{x}_r^T\mathbf{x}_s = 0$. The eigenvalues λ_r and λ_s and the corresponding eigenvectors $\mathbf{x}_r$ and $\mathbf{x}_s$ satisfy the equations

$$\mathbf{A}\mathbf{x}_r = \lambda_r\mathbf{x}_r \quad \text{and} \quad \mathbf{A}\mathbf{x}_s = \lambda_s\mathbf{x}_s,$$

from which, after premultiplication by $\mathbf{x}_s^{\mathrm{T}}$ and $\mathbf{x}_r^{\mathrm{T}}$, respectively, we obtain the two scalar equations

$$\mathbf{x}_s^{\mathrm{T}}\mathbf{A}\mathbf{x}_r = \lambda_r \mathbf{x}_s^{\mathrm{T}}\mathbf{x}_r \quad \text{and} \quad \mathbf{x}_r^{\mathrm{T}}\mathbf{A}\mathbf{x}_s = \lambda_s \mathbf{x}_r^{\mathrm{T}}\mathbf{x}_s.$$

Again, using the fact that the transpose of a scalar leaves it unchanged, we see that the preceding results are identical, so subtracting them we arrive at the condition

$$(\lambda_r - \lambda_s)\mathbf{x}_r^{\mathrm{T}}\mathbf{x}_s = 0.$$

As $\lambda_r \neq \lambda_s$ for $r \neq s$, this is only possible if $\mathbf{x}_r^{\mathrm{T}}\mathbf{x}_s = 0$, so the eigenvectors are mutually orthogonal and the proof is complete. ■

It can be shown that even when some of the eigenvalues of a real symmetric $n \times n$ matrix $\mathbf{A}$ are repeated, the matrix $\mathbf{A}$ will still have n linearly independent eigenvectors, though this result will not be proved here. See, for example, references [2.1], [2.5], [2.8], [2.9], and [2.10].

Orthogonal matrices

orthogonal matrices and rotations

An $n \times n$ real matrix $\mathbf{Q}$ will be said to be an **orthogonal** matrix if

$$\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I} \tag{12}$$

so, if $\mathbf{Q}$ is an orthogonal matrix, it follows that

$$\mathbf{Q}^{\mathrm{T}} = \mathbf{Q}^{-1}.$$

When interpreted geometrically in terms of the cartesian geometry of two or three space dimensions, premultiplication of a linear transformation by an orthogonal matrix corresponds to a pure rotation (or a reflection or both; rotation only if $\det Q = +1$) in space that preserves the lengths between any two points in space, and also the angles between any two straight lines.

A typical geometrical interpretation of a two-dimensional transformation performed by an orthogonal matrix has already been encountered in Section 3.2(c), where the transformation considered was $\mathbf{x}' = \mathbf{R}\mathbf{x}$, with

$$\mathbf{R} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x \\ y \end{bmatrix}, \quad \text{and} \quad \mathbf{x}' = \begin{bmatrix} x' \\ y' \end{bmatrix}.$$

When this transformation was considered in Section 3.2(c), the column vector $\mathbf{x}$ represented a point P in the (x, y)-plane with coordinates (x, y), and $\mathbf{x}'$ represented the same point with coordinates (x', y') in the (x', y')-plane, which was obtained by rotating the O$\{x, y\}$ axes counterclockwise through an angle θ about the origin, as shown in Fig. 4.2.

The transformation (interpreted as a mapping of points) shows that every point in the O$\{x', y'\}$ plane experiences the same rotation through an angle θ about the origin. To show that lengths are preserved, let points P_1 and P_2 have coordinates (x_1, y_1) and (x_2, y_2) in the O$\{x, y\}$ plane and their image points P_1' and P_2' have the coordinates (x_1', y_1') and (x_2', y_2') in the O$\{x', y'\}$ plane. Then the square of the distance d between P_1 and P_2 is given by $d^2 = (x_1 - x_2)^2 + (y_1 - y_2)^2$, and the square

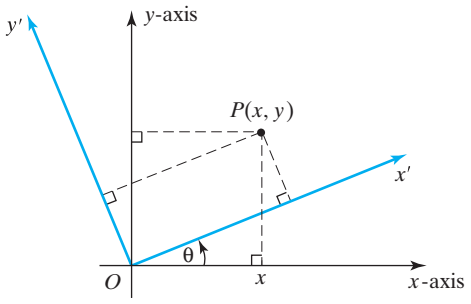


FIGURE 4.2 A rotation of axes about the origin through the angle θ.

of the distance $(d')^2$ between P_1' and P_2' is given by $(d')^2 = (x_1' - x_2')^2 + (y_1' - y_2')^2$. However, from the linear transformation $\mathbf{x}' = \mathbf{R}\mathbf{x}$ we find that

$$x_1 = x_1' \cos\theta - y_1' \sin\theta, \quad x_2 = x_2' \cos\theta - y_2' \sin\theta$$

and

$$y_1 = x_1' \sin\theta + y_1' \cos\theta, \quad y_2 = x_2' \sin\theta + y_2' \cos\theta,$$

from which, after substituting for x_1', x_2', y_1', and y_2', it follows that $(d')^2 = d^2$, showing that distances are preserved. The angles between straight lines in the plane will be preserved because the points on each line will be rotated about the origin through the same angle without changing their distance from the origin.

EXAMPLE 4.8 Show that the matrix

$$\mathbf{R} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

is orthogonal.

Solution We have

$$\mathbf{R}^{\mathrm{T}} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix},$$

but $\mathbf{R}^{\mathrm{T}}\mathbf{R} = \mathbf{I}$, so $\mathbf{R}$ is orthogonal. ■

THEOREM 4.5

main properties of orthogonal matrices

Properties of orthogonal matrices

(i) If $\mathbf{Q}$ is orthogonal then $\det\mathbf{Q} = \pm 1$;
(ii) The product of $n \times n$ orthogonal matrices is an orthogonal matrix;
(iii) The eigenvalues of an orthogonal matrix are all of unit modulus;
(iv) The rows (columns) of an orthogonal matrix form an orthonormal set of vectors.

Proof To prove (i) we start from the fact that $\det\mathbf{Q} = \det\mathbf{Q}^{\mathrm{T}}$. This follows directly from the Laplace expansion of a determinant, because expanding $\det\mathbf{Q}$ in terms of the elements of its ith row is the same as expanding $\det\mathbf{Q}^{\mathrm{T}}$ in terms of the elements of its ith column. From (12), $\mathbf{Q}\mathbf{Q}^{\mathrm{T}} = \mathbf{1}$, so as $\det(\mathbf{AB}) = \det\mathbf{A}\det\mathbf{B}$ we can write $\det\mathbf{Q}\det\mathbf{Q}^{\mathrm{T}} = 1$, but $\det\mathbf{Q}^{\mathrm{T}} = \det\mathbf{Q}$ by Theorem 3.4 so $\det\mathbf{Q}\det\mathbf{Q}^{\mathrm{T}} = (\det\mathbf{Q})^2 = 1$,

and so $\det\mathbf{Q} = \pm 1$. If $\det \mathbf{Q} = +1$, rotation. If $\det \mathbf{Q} = -1$, rotation plus reflection in general.

Result (ii) follows from the fact that if $\mathbf{Q}_1$ and $\mathbf{Q}_2$ are two $n \times n$ orthogonal matrices, then $(\mathbf{Q}_1\mathbf{Q}_2)^{\mathrm{T}}\mathbf{Q}_1\mathbf{Q}_2 = \mathbf{Q}_2^{\mathrm{T}}\mathbf{Q}_1^{\mathrm{T}}\mathbf{Q}_1\mathbf{Q}_2 = \mathbf{Q}_2^{\mathrm{T}}\mathbf{Q}_2 = \mathbf{1}$, and the result is established.

The proof of Result (iii) is similar to the proof of (i) in Theorem 4.3. If $\mathbf{Q}$ is real, taking the complex conjugate of $\mathbf{Qx} = \lambda\mathbf{x}$ gives $\mathbf{Q}\overline{\mathbf{x}} = \overline{\lambda}\overline{\mathbf{x}}$, so taking the transpose of this we find that $\overline{\mathbf{x}}^{\mathrm{T}}\mathbf{Q}^{\mathrm{T}} = \overline{\lambda}\overline{\mathbf{x}}^{\mathrm{T}}$. Forming the product of these two results gives $\overline{\mathbf{x}}^{\mathrm{T}}\mathbf{Q}^{\mathrm{T}}\mathbf{Qx} = \lambda\overline{\lambda}\overline{\mathbf{x}}^{\mathrm{T}}\mathbf{x}$, but $\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I}$, so $\overline{\mathbf{x}}^{\mathrm{T}}\mathbf{x} = \lambda\overline{\lambda}\overline{\mathbf{x}}^{\mathrm{T}}\mathbf{x}$, showing that $\lambda\overline{\lambda} = 1$. Result (iii) follows from this last result because $\lambda\overline{\lambda} = |\lambda|^2 = 1$.

Finally, Result (iv) follows from the definition of an orthogonal matrix, because $\mathbf{QQ}^{\mathrm{T}} = 1$, and if $\mathbf{u}_i$ is the ith row of $\mathbf{Q}$ and $\mathbf{v}_j$ is the jth column of $\mathbf{Q}^{\mathrm{T}}$ (the jth column of $\mathbf{Q}$), then $\mathbf{u}_i\mathbf{v}_j = 0$ for $i \neq j$, and $\mathbf{u}_i\mathbf{v}_j = 1$ for $i = j$, confirming that the vectors form an orthonormal set. ■

Summary

After definition of the eigenvalues of an $n \times n$ matrix **A** in terms of its characteristic polynomial, the associated eigenvectors were defined. An eigenvalue that is repeated r times was said to have the algebraic multiplicity r, and the set of all eigenvalues of **A** was called the spectrum of **A**. The spectral radius of **A** was defined in terms of the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ as the number $R = \max\{|\lambda_1|, |\lambda_2|, \ldots |\lambda_n|\}$, and the linear independence of the set of all eigenvectors was established. The most frequently used method of normalizing eigenvectors was introduced, and examples were worked showing how to determine eigenvectors once the eigenvalues are known.

A simple test was given to check the sum of all eigenvalues, and the Gerschgorin circle theorem was proved that determines a region inside which all eigenvalues must lie, though the region determined in this manner is far from optimal. Inner products, the norm, and systems of orthogonal and orthonormal vectors were introduced, and the most important eigenvalue and eigenvector properties of symmetric matrices and orthogonal matrices were derived.

EXERCISES 4.1

In Exercises 1 through 8, find the characteristic polynomial of the given matrix.

1. $\begin{bmatrix} 2 & 1 & 3 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$.

2. $\begin{bmatrix} 2 & 1 & 3 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}$.

3. $\begin{bmatrix} 1 & 0 & 2 \\ -1 & 1 & -1 \\ 0 & 2 & 1 \end{bmatrix}$.

4. $\begin{bmatrix} 3 & 1 & 1 \\ -2 & 2 & 1 \\ 1 & -1 & 2 \end{bmatrix}$.

5. $\begin{bmatrix} -1 & 0 & 1 \\ 3 & 2 & 1 \\ 1 & 2 & 3 \end{bmatrix}$.

6. $\begin{bmatrix} 4 & 1 & -1 \\ 1 & 0 & 2 \\ -1 & 1 & 2 \end{bmatrix}$.

7. $\begin{bmatrix} 1 & 1 & -1 & 0 \\ 1 & -1 & 1 & 0 \\ 1 & -3 & 3 & 0 \\ -1 & 2 & -1 & -1 \end{bmatrix}$.

8. $\begin{bmatrix} -1 & 1 & 0 & 1 \\ -1 & 2 & -1 & 1 \\ 5 & -3 & 4 & -5 \\ 3 & -2 & 3 & -3 \end{bmatrix}$.

In Exercises 9 through 24 find the eigenvalues and eigenvectors of the given matrix.

9. $\begin{bmatrix} 3 & -2 & 2 \\ 6 & -4 & 6 \\ 2 & -1 & 3 \end{bmatrix}$.

10. $\begin{bmatrix} 3 & -1 & 1 \\ 4 & -1 & 4 \\ 2 & -1 & 4 \end{bmatrix}$.

11. $\begin{bmatrix} -3 & 2 & -2 \\ 4 & -1 & 4 \\ 8 & -4 & 7 \end{bmatrix}$.

12. $\begin{bmatrix} 3 & -2 & 4 \\ -4 & 5 & -4 \\ -4 & 4 & -5 \end{bmatrix}$.

13. $\begin{bmatrix} -5 & 4 & -1 \\ -3 & 2 & -1 \\ 6 & -4 & 2 \end{bmatrix}$.

14. $\begin{bmatrix} 0 & 1 & -2 \\ 2 & -1 & 2 \\ 2 & -2 & 4 \end{bmatrix}$.

15. $\begin{bmatrix} -5 & 8 & 1 \\ -3 & 6 & 1 \\ 6 & -8 & 0 \end{bmatrix}$.

16. $\begin{bmatrix} -1 & 0 & -2 \\ -1 & 2 & -1 \\ 4 & 0 & 5 \end{bmatrix}$.

17. $\begin{bmatrix} -1 & 0 & 2 \\ -1 & 2 & 0 \\ -1 & 0 & 2 \end{bmatrix}$.

18. $\begin{bmatrix} 6 & 0 & 4 \\ 3 & 1 & 3 \\ -8 & 0 & -6 \end{bmatrix}$.

19. $\begin{bmatrix} 0 & 0 & 2 \\ -1 & 1 & 2 \\ -1 & 0 & 3 \end{bmatrix}$.

20. $\begin{bmatrix} 4 & 0 & 2 \\ 2 & 2 & 2 \\ -4 & 0 & 2 \end{bmatrix}$.

21. $\begin{bmatrix} 4 & 0 & -4 \\ 2 & 2 & -4 \\ 2 & 0 & -2 \end{bmatrix}$.

22. $\begin{bmatrix} 3 & 0 & 1 \\ 2 & 1 & 1 \\ -2 & 0 & 0 \end{bmatrix}$.

23. $\begin{bmatrix} -1 & -1 & 1 & 0 \\ 1 & 1 & 1 & -1 \\ 1 & 3 & -1 & -1 \\ -2 & 2 & -2 & 1 \end{bmatrix}$.

24. $\begin{bmatrix} 0 & 1 & 0 & -1 \\ 1 & 0 & 0 & -1 \\ 1 & -2 & 0 & -1 \\ -3 & 3 & 0 & 2 \end{bmatrix}$.

25. Prove that the eigenvalues of upper and lower triangular matrices are equal to the elements on the leading diagonal. Show by example that, unlike the case of diagonal matrices, an eigenvalue of an upper or lower triangular matrix with algebraic multiplicity r has fewer than r eigenvectors.

26. Apply the Gerschgorin circle theorem to one or more of the matrices in Exercises 9 through 24 to verify that the eigenvalues lie within or on the circles determined by the theorem.

27. It can be shown that all the zeros of the polynomial

$$P_n(\lambda) = a_0 + a_1\lambda + a_2\lambda^2 + \cdots + a_n\lambda^n, \quad a_n \neq 0,$$

lie in the circle

$$|\lambda| < 1 + \max\left|\frac{a_k}{a_n}\right|, \quad k = 0, 1, 2, \ldots, n-1.$$

Verify this result by applying it to one or more of the characteristic equations associated with the matrices in Exercises 9 through 24.

The Routh–Hurwitz stability criterion

Let the real polynomial $P_n(\lambda)$ be given by

$$P_n(\lambda) = \lambda^n + a_1\lambda^{n-1} + a_2\lambda^{n-2} + \cdots + a_n$$

and form the determinants

$$\Delta_1 = a_1, \quad \Delta_2 = \begin{vmatrix} a_1 & a_3 \\ 1 & a_2 \end{vmatrix}, \quad \Delta_3 = \begin{vmatrix} a_1 & a_3 & a_5 \\ 1 & a_2 & a_4 \\ 0 & a_1 & a_3 \end{vmatrix}, \ldots,$$

$$\Delta_n = \begin{vmatrix} a_1 & a_3 & a_5 & \ldots & a_{2n-1} \\ 1 & a_2 & a_4 & \ldots & a_{2n-2} \\ 0 & a_1 & a_3 & \ldots & a_{2n-3} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0 & a_n \end{vmatrix} \quad \text{with } a_k = 0 \text{ for } k > n.$$

Then, $\Delta_r > 0$ for $r = 1, 2, \ldots, n$, if and only if every zero of $P_n(\lambda)$ has a negative real part.

28.

(a) Numerical computation shows that the matrix

$$\mathbf{A} = \begin{bmatrix} -2 & 1 & 5 \\ 2 & 3 & 1 \\ 0 & 4 & 2 \end{bmatrix}$$

has the eigenvalues 5.7238, $-1.3619 + 1.9328i$, and $-1.3619 - 1.9328i$. Apply the Routh–Hurwitz stability criterion to confirm that not every zero of the characteristic polynomial has a negative real part.

(b) Numerical computation shows that the matrix

$$\mathbf{A} = \begin{bmatrix} -2 & -2 & -3 \\ 3 & -1 & 0 \\ -4 & 0 & -3 \end{bmatrix}$$

has the eigenvalues -5.4873, $-0.2563 - 1.4564i$, and $-0.2563 + 1.4564i$. Apply the Routh–Hurwitz stability criterion to confirm that every zero of the characteristic polynomial has a negative real part.

An $n \times n$ matrix $\mathbf{A}$ is said to be **similar** to an $n \times n$ matrix $\mathbf{B}$ if there exists a nonsingular $n \times n$ matrix $\mathbf{M}$ such that $\mathbf{B} = \mathbf{M}^{-1}\mathbf{A}\mathbf{M}$. The relationship between $\mathbf{A}$ and $\mathbf{B}$ is said to constitute a **similarity transformation** between the two matrices.

29. If $\mathbf{A}$ and $\mathbf{B}$ are similar, show that $\det\mathbf{A} = \det\mathbf{B}$, and by substituting $\mathbf{B} = \mathbf{M}^{-1}\mathbf{A}\mathbf{M}$ in $\det\mathbf{B}$ and expanding the result, show that similar matrices have the same eigenvalues.

30. Verify the result of Exercise 29 by direct calculation by using

$$\mathbf{A} = \begin{bmatrix} 3 & 1 & -1 \\ 4 & 0 & -1 \\ 4 & -2 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{M} = \begin{bmatrix} 1 & 4 & 1 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \end{bmatrix}$$

to show that both $\mathbf{A}$ and $\mathbf{B}$ have the eigenvalues -1, 2, and 3.

31. Let the $n \times n$ elementary matrix $\mathbf{E}$ be obtained from the unit matrix $\mathbf{I}$ by interchanging its ith and jth rows (columns). By considering the product $\mathbf{E}\mathbf{Q}$, where $\mathbf{Q}$ is an $n \times n$ orthogonal matrix, prove that an orthogonal matrix remains orthogonal when its rows (columns) are interchanged.

4.2 Diagonalization of Matrices

Our purpose in this section will be to examine the possibility of diagonalizing an $n \times n$ matrix $\mathbf{A}$. The reason for this is to try to simplify the structure of $\mathbf{A}$ so that, in some ways, it reflects the simple properties of a diagonal matrix. Diagonalization finds many applications, some of which will be discussed later.

diagonal matrix

Let $\mathbf{D}$ be the general $n \times n$ diagonal matrix

$$\mathbf{D} = \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & \lambda_n \end{bmatrix}. \tag{13}$$

Then, as already seen in Section 4.1, the eigenvalues of $\mathbf{D}$ are the entries $\lambda_1, \lambda_2, \ldots, \lambda_n$ on its leading diagonal, and the corresponding n linearly independent eigenvectors can be taken to be

$$\mathbf{x_1} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \cdot \\ \cdot \\ 0 \end{bmatrix}, \quad \mathbf{x_2} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \cdot \\ \cdot \\ 0 \end{bmatrix}, \ldots, \mathbf{x}_n = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \cdot \\ \cdot \\ 1 \end{bmatrix}. \tag{14}$$

The rule for matrix multiplication shows that

$$\mathbf{D}^m = \begin{bmatrix} \lambda_1^m & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2^m & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & \lambda_n^m \end{bmatrix}, \tag{15}$$

for any positive integer m, so $\mathbf{D}^m$ is easily computed and will have the same set of eigenvectors as $\mathbf{D}$, though its eigenvalues will be $\lambda_1^m, \lambda_2^m, \ldots, \lambda_n^m$.

In addition to these properties, it is obvious that $\det\mathbf{D} = \lambda_1 \cdot \lambda_2 \cdots \lambda_n$, so $\mathbf{D}$ will be nonsingular provided no entry on its leading diagonal is zero. As a result, when $\mathbf{D}$ is nonsingular, the rule for matrix multiplication shows that $\mathbf{D}\mathbf{D}^{-1} = \mathbf{I}$, where

$$\mathbf{D}^{-1} = \begin{bmatrix} 1/\lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & 1/\lambda_2 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & 1/\lambda_n \end{bmatrix}. \tag{16}$$

We now state and prove the fundamental theorem on the diagonalization of $n \times n$ matrices.

THEOREM 4.6

how to diagonalize a matrix

Diagonalization of an $n \times n$ matrix Let the $n \times n$ matrix $\mathbf{A}$ have n eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$, not all of which need be distinct, and let there be n corresponding distinct eigenvectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, so that

$$\mathbf{A}\mathbf{x}_i = \lambda_i \mathbf{x}_i, \quad i = 1, 2, \ldots, n.$$

Define the matrix $\mathbf{P}$ to be the $n \times n$ matrix in which the ith column is the eigenvector $\mathbf{x}_i$, with $i = 1, 2, \ldots, n$, so that in partitioned form $\mathbf{P} = [\mathbf{x}_1 \ \mathbf{x}_2 \ \cdots \ \mathbf{x}_n]$, and let $\mathbf{D}$ be the diagonal matrix

$$\mathbf{D} = \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & \lambda_n \end{bmatrix},$$

where the eigenvalue λ_i is in the ith position in the ith row. Then

$$\mathbf{P}^{-1}\mathbf{AP} = \mathbf{D}.$$

Proof Consider the product $\mathbf{B} = \mathbf{AP}$. Then, by expressing $\mathbf{P}$ in partitioned form, we can write $\mathbf{B}$ as

$$\mathbf{B} = [\mathbf{Ax}_1 \quad \mathbf{Ax}_2 \quad \ldots \quad \mathbf{Ax}_n].$$

Using the fact that $\mathbf{Ax}_i = \lambda_i \mathbf{x}_i$ allows this to be rewritten as

$$\mathbf{B} = [\lambda_1\mathbf{x}_1 \quad \lambda_2\mathbf{x}_2 \quad \ldots \quad \lambda_n\mathbf{x}_n] = \mathbf{PD},$$

showing that

$$\mathbf{PD} = \mathbf{AP}.$$

As the columns of $\mathbf{P}$ are linearly independent, $\mathbf{P}$ is nonsingular, so $\mathbf{P}^{-1}$ exists and we can premultiply by $\mathbf{P}^{-1}$ to obtain

$$\mathbf{D} = \mathbf{P}^{-1}\mathbf{AP},$$

and the theorem is proved. ■

General Remarks About Diagonalization

(i) An $n \times n$ matrix can be diagonalized provided it possesses n linearly independent eigenvectors.

(ii) A symmetric matrix can always be diagonalized.

(iii) The diagonalizing matrix for a real $n \times n$ matrix $\mathbf{A}$ may contain complex elements. This is because although the characteristic polynomial of $\mathbf{A}$ has real coefficients, its zeros either will be real or will occur in complex conjugate pairs.

(iv) A diagonalizing matrix is not unique, because its form depends on the order in which the eigenvectors of $\mathbf{A}$ are used to form its columns.

A useful consequence of the diagonalized form of a matrix is that it enables it to be raised to a positive integral power with the minimum of effort. This property will be used later when the matrix exponential is introduced.

To see the ease with which an $n \times n$ matrix can be raised to a power when it is diagonalizable, we start by writing $\mathbf{A}$ in the form $\mathbf{A} = \mathbf{PDP}^{-1}$. We then have

$$\mathbf{A}^2 = (\mathbf{PDP}^{-1})(\mathbf{PDP}^{-1}) = \mathbf{PDP}^{-1}\mathbf{PDP}^{-1} = \mathbf{PDDP}^{-1} = \mathbf{PD}^2\mathbf{P}^{-1},$$

so that, in general,

$$\mathbf{A}^m = \mathbf{P}\mathbf{D}^m\mathbf{P}^{-1}, \quad \text{for } m = 1, 2, \ldots.$$

As evaluating $\mathbf{D}^m$ simply involves raising each entry on its leading diagonal to the power m, the evaluation of $\mathbf{A}^m$ only involves three matrix multiplications.

This last result was used without justification in Section 3.2(f) when a stochastic matrix was raised to the power m (do not confuse the stochastic matrix $\mathbf{P}$ in that section with the orthogonalizing matrix $\mathbf{P}$ just defined).

EXAMPLE 4.9 Diagonalize the matrix

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & -1 \\ 3 & 2 & -3 \\ 3 & 1 & -2 \end{bmatrix},$$

and use the result to find $\mathbf{A}^5$.

Solution Matrix $\mathbf{A}$ was examined in Example 4.1 and shown to have the eigenvalues $\lambda_1 = 2$, $\lambda_2 = 1$, and $\lambda_3 = -1$, and the corresponding eigenvectors

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \quad \text{and} \quad \mathbf{x}_3 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}.$$

Theorem 4.5 shows that a diagonalizing matrix $\mathbf{P}$ is given by

$$\mathbf{P} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix},$$

and a routine calculation shows that

$$\mathbf{P}^{-1} = \begin{bmatrix} 1 & 1 & -1 \\ 0 & -1 & 1 \\ -1 & 0 & 1 \end{bmatrix}.$$

Before finding $\mathbf{A}^5$, and although it is unnecessary for what is to follow, it is instructive to check that when the matrix $\mathbf{P}^{-1}\mathbf{A}\mathbf{P}$ is formed, the eigenvalues appearing in the diagonal matrix $\mathbf{D}$ do so in the order in which the corresponding eigenvectors of $\mathbf{A}$ have been used to form the columns of $\mathbf{P}$. This is seen to be so in this case because

$$\mathbf{D} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}.$$

Returning to the calculation of $\mathbf{A}^5$ and using the expressions for $\mathbf{P}$, $\mathbf{P}^{-1}$, and $\mathbf{D}$ in $\mathbf{A}^5 = \mathbf{P}\mathbf{D}^5\mathbf{P}^{-1}$ gives

$$\mathbf{A}^5 = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 2^5 & 0 & 0 \\ 0 & 1^5 & 0 \\ 0 & 0 & (-1)^5 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 32 & 31 & -31 \\ 33 & 32 & -33 \\ 33 & 31 & -32 \end{bmatrix}.$$

Had the eigenvectors been arranged in a different order when constructing $\mathbf{P}$, a different but equivalent diagonal matrix would have been obtained. For example,

if $\mathbf{P}$ had been written

$$\mathbf{P} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix},$$

$\mathbf{D}$ would have become

$$\mathbf{D} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & -1 \end{bmatrix},$$

though after $\mathbf{P}^{-1}$ was found and $\mathbf{A}^5 = \mathbf{P}\mathbf{D}^5\mathbf{P}^{-1}$ was computed, the matrix $\mathbf{A}^5$ would, of course, remain the same. ■

EXAMPLE 4.10 Diagonalize the matrix

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 1 & 1 \\ -1 & 2 & 0 & 1 \\ -1 & 0 & 2 & 1 \\ 1 & 0 & -1 & 0 \end{bmatrix}.$$

Solution Matrix $\mathbf{A}$ was considered in Example 4.2, which showed that it had the eigenvalues $\lambda_1 = 0$, $\lambda_2 = 1$, $\lambda_3 = 1$, and $\lambda_4 = 2$, and that although the eigenvalue 1 occurred with algebraic multiplicity 2, the matrix still had the four linearly independent eigenvectors

$$(\lambda_1 = 0) \quad \mathbf{x}_1 = \begin{bmatrix} -1 \\ -1 \\ -1 \\ 1 \end{bmatrix}, \quad (\lambda_2 = 1) \quad \mathbf{x}_2 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \end{bmatrix}, \quad (\lambda_3 = 1) \quad \mathbf{x}_3 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix},$$

and

$$(\lambda_4 = 2) \quad \mathbf{x}_1 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}.$$

Using these eigenvectors to form $\mathbf{P}$ gives

$$\mathbf{P} = \begin{bmatrix} -1 & 1 & 1 & 0 \\ -1 & 1 & 0 & 1 \\ -1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{bmatrix},$$

from which it follows that

$$\mathbf{P}^{-1} = \begin{bmatrix} -1 & 0 & 1 & 1 \\ -1 & 0 & 2 & 1 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & -1 & 0 \end{bmatrix}.$$

Because of the ordering of the eigenvectors, the diagonal matrix **D** will be

$$\mathbf{D} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix},$$

where

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{D}.$$

■

We saw in Theorem 4.4 that a real symmetric $n \times n$ matrix **A** with distinct eigenvalues has a set of n mutually orthogonal linearly independent eigenvectors. It follows at once that if when constructing the diagonalizing matrix for **A** the normalized eigenvectors of **A** are used to form the columns of **P**, the resulting diagonalizing matrix will be an *orthogonal* matrix. This is often advantageous, because the properties of orthogonal matrices can simplify subsequent calculations that may arise. However, if an eigenvalue is repeated, the corresponding eigenvectors will not, in general, be orthogonal to the other eigenvectors, so although there will still be a set of n linearly independent eigenvectors, the set will no longer form an orthogonal set.

Because of the frequency with which symmetric matrices arise in applications, and the fact that symmetric matrices with repeated eigenvalues are not unusual, it is reasonable to ask if it is possible for symmetric matrices always to be diagonalized by an orthogonal matrix and, if so, how this can be achieved. The answer to the question about the possibility of diagonalization by an orthogonal matrix is in the affirmative. The method of arriving at an orthonormal set of vectors to be used when constructing **P** involves using a generalization of the Gram–Schmidt orthogonalization process introduced in Section 2.7 in the context of geometrical vectors in R^3.

As an n element matrix vector is simply a vector in a vector space, an extension of the Gram–Schmidt orthogonalization process to include n-element matrix vectors can be used to construct an *orthonormal* set of n vectors from any set of n linearly independent eigenvectors that are always associated with an $n \times n$ symmetric matrix **A**. The required generalization of the orthogonalization process that leads to an **orthonormal system** is an immediate extension of the one derived in Section 2.7, so the details of its derivation will be omitted.

orthogonalization of a set of linearly independent vectors

Rule for the Gram–Schmidt orthogonalization process for matrix vectors

Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ be a set of n element linearly independent nonorthogonal matrix column vectors. Then an equivalent **orthonormal set** of vectors $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n$ can be constructed from the vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, via an intermediate set of orthogonal nonnormalized vectors $\mathbf{v}_2, \mathbf{v}_2, \ldots, \mathbf{v}_n$. The steps involved in the determination of the vectors $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n$ are as follows:

$$\begin{aligned}
\mathbf{p}_1 &= \mathbf{x}_1/\|\mathbf{x}_1\|, \\
\mathbf{v}_2 &= \mathbf{x}_2 - (\mathbf{p}_1 \cdot \mathbf{x}_2)\mathbf{p}_1, \\
\mathbf{p}_2 &= \mathbf{v}_2/\|\mathbf{v}_2\|, \\
\mathbf{v}_r &= \mathbf{x}_r - \{(\mathbf{p}_1 \cdot \mathbf{x}_r)\mathbf{p}_1 + (\mathbf{p}_2 \cdot \mathbf{x}_r)\mathbf{p}_2 + \cdots + (\mathbf{p}_{r-1} \cdot \mathbf{x}_r)\mathbf{p}_{r-1}\} \\
\mathbf{p}_r &= \mathbf{v}_r/\|\mathbf{v}_r\|, \quad \text{for } r = 2, 3, \ldots, n.
\end{aligned}$$

When the Gram–Schmidt orthogonalization process is applied to the eigenvectors of a real symmetric matrix $\mathbf{A}$ with repeated eigenvalues, the diagonalizing matrix $\mathbf{P}$ is constructed by using the vectors $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n$, obtained from the preceding scheme after starting with any linearly independent set of eigenvectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ of $\mathbf{A}$. Then, in partitioned form,

$$\mathbf{P} = [\mathbf{p}_1 \quad \mathbf{p}_2 \quad \ldots \quad \mathbf{p}_n]$$

and, as before,

$$\mathbf{D} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P},$$

where $\mathbf{D}$ is again a diagonal matrix with its diagonal elements equal to the eigenvalues of $\mathbf{A}$ arranged in the same order as the corresponding columns of $\mathbf{P}$. This time, however, entries on the leading diagonal will be repeated as many times as the multiplicity of the eigenvalues concerned.

EXAMPLE 4.11

Use the Gram–Schmidt orthogonalization process to construct an orthonormal set of vectors from the vectors

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}, \quad \text{and} \quad \mathbf{x}_3 = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}.$$

Solution In this case the Gram–Schmidt orthogonalization process involves the three vectors $\mathbf{x}_1, \mathbf{x}_2$, and $\mathbf{x}_3$, so a set of orthonormal vectors $\mathbf{p}_1, \mathbf{p}_2$, and $\mathbf{p}_3$ is given by the scheme

$$\begin{aligned}
\mathbf{p}_1 &= \mathbf{x}_1/\|\mathbf{x}_1\| \\
\mathbf{v}_2 &= \mathbf{x}_2 - (\mathbf{p}_1 \cdot \mathbf{x}_2)\mathbf{p}_1 \\
\mathbf{p}_2 &= \mathbf{v}_2/\|\mathbf{v}_2\| \\
\mathbf{v}_3 &= \mathbf{x}_3 - \{(\mathbf{p}_1 \cdot \mathbf{x}_3)\mathbf{p}_1 + (\mathbf{p}_2 \cdot \mathbf{x}_3)\mathbf{p}_2\} \\
\mathbf{p}_3 &= \mathbf{v}_3/\|\mathbf{v}_3\|.
\end{aligned}$$

A series of straightforward calculations gives

$$\mathbf{p}_1 = \begin{bmatrix} 1/\sqrt{3} \\ 1/\sqrt{3} \\ 1/\sqrt{3} \end{bmatrix}, \quad \text{and} \quad \mathbf{v}_2 = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} - 0\,\mathbf{p}_1 = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}, \quad \text{so } \mathbf{p}_2 = \begin{bmatrix} 1/\sqrt{2} \\ 0 \\ -1/\sqrt{2} \end{bmatrix},$$

and, finally,

$$\mathbf{v}_3 = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix} - \sqrt{3}\begin{bmatrix} 1/\sqrt{3} \\ 1/\sqrt{3} \\ 1/\sqrt{3} \end{bmatrix} - 1/\sqrt{2}\begin{bmatrix} 1/\sqrt{2} \\ 0 \\ -1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} -1/2 \\ 1 \\ -1/2 \end{bmatrix},$$

so

$$\mathbf{p}_3 = \begin{bmatrix} -1/\sqrt{6} \\ \sqrt{(2/3)} \\ -1/\sqrt{6} \end{bmatrix}.$$

■

EXAMPLE 4.12 Construct an orthogonal diagonalizing matrix for the symmetric matrix

$$\mathbf{A} = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 1 & 2 \\ 0 & 2 & 1 \end{bmatrix}.$$

Solution This has the *distinct* eigenvalues $\lambda_1 = -1$, $\lambda_2 = 3$, and $\lambda_1 = 4$, so the corresponding eigenvectors $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$ are orthogonal. Simple calculations show that

$$\mathbf{x}_1 = \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}, \quad \text{and} \quad \mathbf{x}_3 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

The normalized eigenvectors are

$$\hat{\mathbf{x}}_1 = \begin{bmatrix} 0 \\ -1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}, \quad \hat{\mathbf{x}}_2 = \begin{bmatrix} 0 \\ 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}, \quad \text{and} \quad \hat{\mathbf{x}}_3 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix},$$

so the diagonalizing matrix $\mathbf{P}$ and the corresponding diagonal matrix $\mathbf{D}$ are

$$\mathbf{P} = \begin{bmatrix} 0 & 0 & 1 \\ -1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 1/\sqrt{2} & 1/\sqrt{2} & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{D} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 4 \end{bmatrix}. \qquad \blacksquare$$

EXAMPLE 4.13 Construct an orthogonal diagonalizing matrix for the real symmetric matrix

$$\mathbf{A} = \begin{bmatrix} -1 & 2 & 4 \\ 2 & 2 & -2 \\ 4 & -2 & -1 \end{bmatrix}.$$

Solution This has the eigenvalues $\lambda_1 = -6$, $\lambda_2 = 3$, and $\lambda_3 = 3$, so as the eigenvalue 3 has multiplicity 2, the corresponding set of eigenvectors $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$ will *not* be orthogonal. The eigenvectors $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$ are easily shown to be

$$\mathbf{x}_1 = \begin{bmatrix} -2 \\ 1 \\ 2 \end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}, \quad \text{and} \quad \mathbf{x}_3 = \begin{bmatrix} 0 \\ -2 \\ 1 \end{bmatrix}.$$

Applying the Gram–Schmidt orthogonalization process to vectors $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$, as in Example 4.11, after some straightforward calculations we arrive at the orthonormal set

$$\mathbf{p}_1 = \begin{bmatrix} -2/3 \\ 1/3 \\ 2/3 \end{bmatrix}, \quad \mathbf{p}_2 = \begin{bmatrix} 1/\sqrt{5} \\ 2/\sqrt{5} \\ 0 \end{bmatrix}, \quad \text{and} \quad \mathbf{p}_3 = \begin{bmatrix} 4/(3\sqrt{5}) \\ -2/(3\sqrt{5}) \\ \sqrt{5}/3 \end{bmatrix}.$$

In this case an orthogonal diagonalizing matrix is

$$\mathbf{P} = \begin{bmatrix} -2/3 & 1/\sqrt{5} & 4/(3\sqrt{5}) \\ 1/3 & 2/\sqrt{5} & -2/(3\sqrt{5}) \\ 2/3 & 0 & \sqrt{5}/3 \end{bmatrix},$$

and the corresponding diagonal matrix is

$$\mathbf{D} = \begin{bmatrix} -6 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 3 \end{bmatrix}.$$

■

To close this section we state the important Cayley–Hamilton theorem, which is true for *all* square matrices, though before considering the theorem we first define a matrix polynomial.

A **matrix polynomial** involving an $n \times n$ matrix $\mathbf{A}$ is an expression of the form

$$\mathbf{A}^m + b_1\mathbf{A}^{m-1} + b_2\mathbf{A}^{m-2} + \cdots + b_{m-1}\mathbf{A} + b_m\mathbf{I},$$

in which m is an integer and $b_1, b_2, \ldots, b_m$ are real or complex numbers.

THEOREM 4.7

a matrix satisfies its own characteristic equation

The Cayley–Hamilton theorem Let $P_n(\lambda)$ be the characteristic polynomial of an arbitrary $n \times n$ square matrix $\mathbf{A}$. Then $\mathbf{A}$ satisfies its own characteristic equation, and so is a solution of the matrix polynomial equation $P_n(\mathbf{A}) = \mathbf{0}$.

Proof For simplicity, we only prove the theorem for real symmetric matrices, though it is true for every $n \times n$ matrix. If $\mathbf{A}$ is a real $n \times n$ symmetric matrix, then from Theorem 4.6 we may write $\mathbf{A} = \mathbf{PDP}^{-1}$. Let the characteristic polynomial of $\mathbf{A}$ be

$$P_n(\lambda) = (-1)^n\{\lambda^n + c_1\lambda^{n-1} + \cdots + c_{n-1}\lambda + c_n\}.$$

Then replacing λ by $\mathbf{A}$ converts $P_n(\lambda)$ to the matrix polynomial

$$P_n(\mathbf{A}) = (-1)^n\{\mathbf{A}^n + c_1\mathbf{A}^{n-1} + \cdots + c_{n-1}\mathbf{A} + c_n\mathbf{I}\},$$

but $\mathbf{A}^r = \mathbf{PD}^r\mathbf{P}^{-1}$, so

$$P_n(\mathbf{A}) = (-1)^n\{\mathbf{P}\{\mathbf{D}^n + c_1\mathbf{D}^{n-1} + \cdots + c_{n-1}\mathbf{D} + c_n\mathbf{I}_n\}\mathbf{P}^{-1}\}.$$

The ith row of the matrix polynomial $\mathbf{D}^n + c_1\mathbf{D}^{n-1} + \cdots + c_{n-1}\mathbf{D} + c_n\mathbf{I}$ is simply $\lambda_i^n + c_1\lambda_i^{n-1} + \cdots + c_{n-1}\lambda_i + c_n$, but this is $P_n(\lambda_i)$, and it must vanish for $i = 1, 2, \ldots, n$ because λ_i is an eigenvalue of $\mathbf{A}$. Thus, $\mathbf{D}^n + c_1\mathbf{D}^{n-1} + \cdots + c_{n-1}\mathbf{D} + c_n\mathbf{I} = \mathbf{0}$, showing that $P_n(\mathbf{A}) = \mathbf{P}\{\mathbf{0}\}\mathbf{P}^{-1} = \mathbf{0}$, and the result is proved. ■

EXAMPLE 4.14

Verify the Cayley–Hamilton theorem for the matrix

$$\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 5 & 2 \end{bmatrix}.$$

Solution The characteristic polynomial is $P_2(\lambda) = \lambda^2 - 4\lambda - 1$, and

$$\mathbf{A}^2 = \begin{bmatrix} 9 & 4 \\ 20 & 9 \end{bmatrix}, \quad \text{so } P_2(\mathbf{A}) = \begin{bmatrix} 9 & 4 \\ 20 & 9 \end{bmatrix} - 4\begin{bmatrix} 2 & 1 \\ 5 & 2 \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

■

Finding $\mathbf{A}^{-1}$ from the Cayley–Hamilton theorem

If the $n \times n$ matrix $\mathbf{A}$ is nonsingular, the following interesting result can be obtained directly from the Cayley–Hamilton theorem. Let the characteristic

polynomial of $\mathbf{A}$ be $P_n(\lambda) = (-1)^n\{\lambda^n + c_1\lambda^{n-1} + \cdots + c_{n-1}\lambda + c_n\}$, so from Theorem 4.7

$$\mathbf{A}^n + c_1\mathbf{A}^{n-1} + \cdots + c_{n-1}\mathbf{A} + c_n\mathbf{I} = 0.$$

The matrix $\mathbf{A}^{-1}$ exists because by hypothesis $\mathbf{A}$ is nonsingular, so premultiplication of the preceding equation by $\mathbf{A}^{-1}$, followed by a rearrangement of terms, allows $\mathbf{A}^{-1}$ to be expressed in terms of powers of $\mathbf{A}$ through the result

$$\mathbf{A}^{-1} = (-1/c_n)\{\mathbf{A}^{n-1} + c_1\mathbf{A}^{n-2} + \cdots + c_{n-1}\mathbf{I}\}. \tag{17}$$

EXAMPLE 4.15 Use the result of equation (17) to find $\mathbf{A}^{-1}$ for the nonsingular matrix

$$\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 5 & 2 \end{bmatrix}.$$

Solution Matrix $\mathbf{A}$ was considered in Example 4.14, where it was found that the characteristic polynomial $P_2(\lambda) = \lambda^2 - 4\lambda - 1$, so in terms of (17) we see that $c_1 = -4$ and $c_2 = -1$. Thus,

$$\mathbf{A}^{-1} = -1/(-1)\left\{\begin{bmatrix} 2 & 1 \\ 5 & 2 \end{bmatrix} - 4\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right\} = \begin{bmatrix} -2 & 1 \\ 5 & -2 \end{bmatrix}. \quad \blacksquare$$

Summary

This section has described how an $n \times n$ matrix can be diagonalized when it possesses n linearly independent eigenvectors. The diagonalization was shown not to be unique, since its form depends on the order in which the eigenvectors are used to construct the diagonalizing matrix $\mathbf{P}$.

Sometimes, when a linearly independent set of n vectors has been obtained, it is desirable to replace it by an equivalent set of n orthogonal or orthonormal vectors. The section closed by showing how this can be accomplished by means of the Gram–Schmidt orthogonalization procedure.

EXERCISES 4.2

In Exercises 1 through 12, find a diagonalizing matrix $\mathbf{P}$ for the given matrix, in each case using the fact that the zeros of the characteristic polynomial are small integers that can be found by trial and error.

1. $\begin{bmatrix} -2 & -3 & -1 \\ 1 & 2 & 1 \\ 3 & 3 & 2 \end{bmatrix}$.

2. $\begin{bmatrix} 3 & 1 & 4 \\ -4 & -2 & -4 \\ -1 & -1 & 2 \end{bmatrix}$.

3. $\begin{bmatrix} 3 & 1 & -2 \\ 6 & 2 & -6 \\ 4 & 1 & -3 \end{bmatrix}$.

4. $\begin{bmatrix} -6 & -10 & -4 \\ 2 & 3 & 2 \\ 7 & 10 & 5 \end{bmatrix}$.

5. $\begin{bmatrix} -1 & 2 & -2 \\ 2 & -1 & 2 \\ 2 & -2 & 3 \end{bmatrix}$.

6. $\begin{bmatrix} 14 & 2 & 8 \\ -8 & -3 & -4 \\ -26 & -4 & -15 \end{bmatrix}$.

7. $\begin{bmatrix} 5 & -2 & 2 \\ 2 & 1 & 2 \\ -2 & 2 & 1 \end{bmatrix}$.

8. $\begin{bmatrix} 12 & 4 & 6 \\ -6 & -2 & -3 \\ -22 & -8 & -11 \end{bmatrix}$.

9. $\begin{bmatrix} 2 & 0 & 0 \\ 1 & -1 & 2 \\ -2 & 0 & 1 \end{bmatrix}$.

10. $\begin{bmatrix} 12 & -4 & 8 \\ -6 & 2 & -4 \\ -20 & 8 & -14 \end{bmatrix}$.

11. $\begin{bmatrix} -6 & 2 & -4 \\ -4 & 0 & -4 \\ 4 & -2 & 2 \end{bmatrix}$.

12. $\begin{bmatrix} -7 & 0 & -6 \\ 3 & -1 & 3 \\ 9 & 0 & 8 \end{bmatrix}$.

In Exercises 13 through 16 use the Gram–Schmidt orthogonalization process with the given set of vectors to find (a) an equivalent set of orthogonal vectors and (b) an orthonormal set.

13. $\begin{bmatrix}1\\1\\1\end{bmatrix}, \begin{bmatrix}0\\1\\1\end{bmatrix}, \begin{bmatrix}0\\0\\1\end{bmatrix}$.

14. $\begin{bmatrix}2\\1\\1\end{bmatrix}, \begin{bmatrix}1\\-1\\1\end{bmatrix}, \begin{bmatrix}0\\2\\1\end{bmatrix}$.

15. $\begin{bmatrix}-1\\1\\0\end{bmatrix}, \begin{bmatrix}2\\1\\-1\end{bmatrix}, \begin{bmatrix}1\\-2\\2\end{bmatrix}$.

16. $\begin{bmatrix}-1\\2\\0\end{bmatrix}, \begin{bmatrix}1\\1\\-1\end{bmatrix}, \begin{bmatrix}1\\-1\\1\end{bmatrix}$.

In Exercises 17 through 22 find an orthogonal diagonalizing matrix **P** for the given symmetric matrix.

17. $\begin{bmatrix}3&0&0\\0&3&1\\0&1&3\end{bmatrix}$.

18. $\begin{bmatrix}5&1&0\\1&5&0\\0&0&2\end{bmatrix}$.

19. $\begin{bmatrix}4&1&0\\1&4&0\\0&0&3\end{bmatrix}$.

20. $\begin{bmatrix}2&1&1\\1&2&1\\1&1&2\end{bmatrix}$.

21. $\begin{bmatrix}4&2&0\\2&4&0\\0&0&2\end{bmatrix}$.

22. $\begin{bmatrix}4&1&1\\1&4&1\\1&1&4\end{bmatrix}$.

23. Verify by direct calculation that the matrix in Exercise 1 satisfies the Cayley–Hamilton theorem.

24. Verify by direct calculation that the matrix in Exercise 7 satisfies the Cayley–Hamilton theorem.

In Exercises 25 through 28 use (17) to find $\mathbf{A}^{-1}$ and check the result by showing that $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$.

25. $\mathbf{A} = \begin{bmatrix}2&3\\-1&4\end{bmatrix}$.

26. $\mathbf{A} = \begin{bmatrix}5&1\\3&-2\end{bmatrix}$.

27. $\mathbf{A} = \begin{bmatrix}2&1&0\\-2&1&2\\0&-1&-2\end{bmatrix}$.

28. $\mathbf{A} = \begin{bmatrix}1&0&2\\3&1&0\\0&2&4\end{bmatrix}$.

4.3 Special Matrices with Complex Elements

In the previous section it was seen that one way in which matrices with complex elements can occur is when the eigenvectors of an arbitrary $n \times n$ matrix are used to construct a diagonalizing matrix. This is not the only reason for considering $n \times n$ matrices with complex elements, because the following three special types of matrices arise naturally in applications of mathematics to physics and engineering, and elsewhere.

Hermitian, skew-Hermitian, and unitary matrices

Let $\mathbf{A} = [a_{ij}]$ be an $n \times n$ matrix with possibly complex elements. Then:

$\mathbf{A}$ is called an **Hermitian** matrix if $\overline{\mathbf{A}}^{\mathrm{T}} = \mathbf{A}$, so that $\overline{a}_{kj} = a_{jk}$;

$\mathbf{A}$ is called a **skew-Hermitian** matrix if $\overline{\mathbf{A}}^{\mathrm{T}} = -\mathbf{A}$, so that $\overline{a}_{kj} = -a_{jk}$;

$\mathbf{U}$ is called a **unitary** matrix if $\overline{\mathbf{U}}^{\mathrm{T}} = \mathbf{U}^{-1}$.

The basic properties of these three types of matrices follow almost directly from their definitions.

Basic Properties of Hermitian, Skew-Hermitian, and Unitary Matrices

1. The elements on the leading diagonal of an Hermitian matrix are real, because $\overline{a}_{ii} = a_{ii}$, and this is only possible if a_{ii} is real.
2. The elements on the leading diagonal of a skew-Hermitian matrix are either purely imaginary or 0. This follows from the fact that $\overline{a}_{ii} = -a_{ii}$, so the real part of a_{ii} must equal its negative, and this is only possible if a_{ii} is purely imaginary or 0.

3. If the elements of an Hermitian matrix are real, then the matrix is a real symmetric matrix, because then $\overline{\mathbf{A}}^{\mathrm{T}} = \mathbf{A}^{\mathrm{T}}$, and the definition of an Hermitian matrix reduces to the definition of a real symmetric matrix.
4. If the elements of a skew-Hermitian matrix are real, then the matrix is a skew-symmetric matrix, because then the definition of a skew-Hermitian matrix reduces to the definition of a skew-symmetric matrix.
5. Any $n \times n$ matrix $\mathbf{A}$ of the form $\mathbf{A} = \mathbf{B} + i\mathbf{C}$, where $\mathbf{B}$ is a real symmetric matrix and $\mathbf{C}$ is a real skew-symmetric matrix, is an Hermitian matrix. This follows directly from Properties 3 and 4.
6. Any $n \times n$ matrix $\mathbf{A}$ can be written in the form $\mathbf{A} = \mathbf{B} + \mathbf{C}$, where $\mathbf{B}$ is Hermitian and $\mathbf{C}$ is a skew-Hermitian. To see this we write $\mathbf{A} = (1/2)(\mathbf{A} + \overline{\mathbf{A}}^{\mathrm{T}}) + (1/2)(\mathbf{A} - \overline{\mathbf{A}}^{\mathrm{T}})$, and then set $\mathbf{B} = (1/2)(\mathbf{A} + \overline{\mathbf{A}}^{\mathrm{T}})$ and $\mathbf{C} = (1/2)(\mathbf{A} - \overline{\mathbf{A}}^{\mathrm{T}})$. Then $\overline{\mathbf{B}}^{\mathrm{T}} = (1/2)(\overline{\mathbf{A}^{\mathrm{T}} + \overline{\mathbf{A}}}) = (1/2)(\mathbf{A} + \overline{\mathbf{A}}^{\mathrm{T}}) = \mathbf{B}$ and $\mathbf{C}^{\mathrm{T}} = (1/2)(\overline{\mathbf{A}^{\mathrm{T}} - \overline{\mathbf{A}}}) = -(1/2)(\mathbf{A} - \overline{\mathbf{A}}^{\mathrm{T}}) = -\mathbf{C}$, showing that $\mathbf{B}$ is Hermitian and $\mathbf{C}$ is skew-Hermitian.
7. A real unitary matrix is an orthogonal matrix, because in that case $\overline{\mathbf{A}}^{\mathrm{T}} = \mathbf{A}^{\mathrm{T}}$, causing the definition of a unitary matrix to reduce to the definition of an orthogonal matrix.
8. The determinant of a unitary matrix is ± 1. This result is established in essentially the same way as the result of Theorem 4.4(i), so the argument will not be repeated.

EXAMPLE 4.16 The following are examples of Hermitian, skew-Hermitian, and unitary matrices.

Hermitian matrix:

$$\mathbf{A} = \begin{bmatrix} 3 & 2+5i & -7+3i \\ 2-5i & 0 & 1-i \\ -7-3i & 1+i & 4 \end{bmatrix}.$$

Skew-Hermitian matrix:

$$\mathbf{B} = \begin{bmatrix} 4i & -3-2i & -6-4i \\ 3-2i & -2i & 5 \\ 6-4i & -5 & 0 \end{bmatrix}.$$

Unitary matrix:

$$\mathbf{U} = \begin{bmatrix} \dfrac{1+i}{2} & \dfrac{-1+i}{2} & 0 \\ \dfrac{1+i}{2} & \dfrac{1-i}{2} & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

■

It can be seen from Properties 3, 4, and 7 that Hermitian, skew-Hermitian, and unitary matrices are, respectively, generalizations of symmetric, skew-symmetric, and orthogonal real-valued matrices. Accordingly, it is to be expected that some of the properties exhibited by these real-valued matrices are shared by their complex generalizations, and this is indeed the case as we now show.

THEOREM 4.8 **Eigenvalues of Hermitian, skew-Hermitian, and unitary matrices**

(i) The eigenvalues of an Hermitian matrix are real.

(ii) The eigenvalues of a skew-Hermitian matrix are either purely imaginary or 0.

(iii) The eigenvalues λ of a unitary matrix are all such that $|\lambda| = 1$.

Proof

(i) Apart for the need to introduce the complex conjugate operation, the proof is essentially the same as that of Theorem 4.4 for symmetric matrices, and so it is omitted.

(ii) Let $\mathbf{x}$ be the eigenvector of $\mathbf{A}$ corresponding to the eigenvalue λ, so $\mathbf{Ax} = \lambda\mathbf{x}$. Then $\overline{\mathbf{x}}^{\mathrm{T}}\mathbf{Ax} = \lambda\overline{\mathbf{x}}^{\mathrm{T}}\mathbf{x}$, from which we have

$$\lambda = \overline{\mathbf{x}}^{\mathrm{T}}\mathbf{Ax}/\overline{\mathbf{x}}^{\mathrm{T}}\mathbf{x},$$

but $\overline{\mathbf{x}}^{\mathrm{T}}\mathbf{x} = x_1\overline{x}_1 + x_2\overline{x}_2 + \cdots + x_n\overline{x}_n$ is real. However, $\overline{\mathbf{A}} = -\mathbf{A}^{\mathrm{T}}$, so $\overline{\mathbf{x}}^{\mathrm{T}}\mathbf{Ax} = -\overline{\mathbf{x}}^{\mathrm{T}}\mathbf{Ax}$, so we can write

$$\lambda = \overline{\mathbf{x}}^{\mathrm{T}}\mathbf{Ax}/\overline{\mathbf{x}}^{\mathrm{T}}\mathbf{x} = -\overline{\overline{\mathbf{x}}^{\mathrm{T}}\mathbf{Ax}/\overline{\mathbf{x}}^{\mathrm{T}}\mathbf{x}}.$$

The product $\overline{\mathbf{x}}^{\mathrm{T}}\mathbf{x}$ is real, so this last result shows that the complex number λ equals the negative of its complex conjugate, and this is only possible if λ is purely imaginary or 0, so the proof is complete.

(iii) Apart from the need to introduce the complex conjugate operation, the proof is essentially that of Theorem 4.5(iii), so it will be omitted. ■

The location of the eigenvalues of these complex matrices and of their corresponding real forms are illustrated in Fig. 4.3.

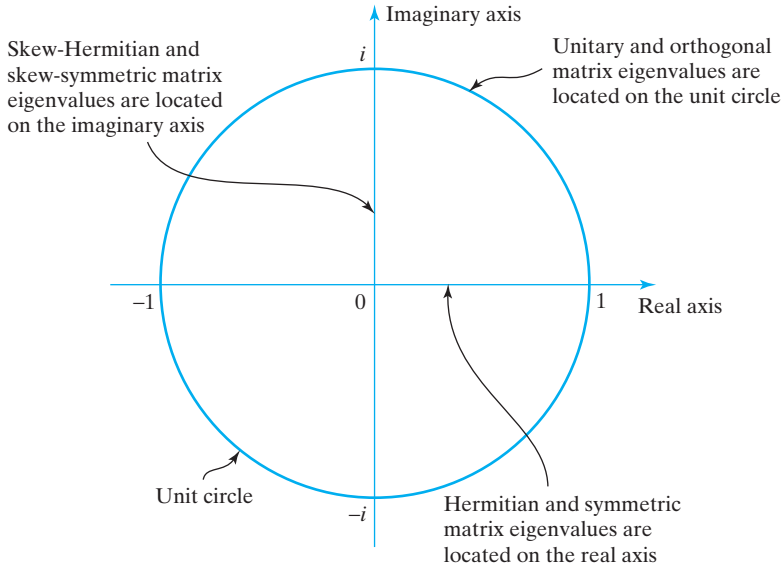


FIGURE 4.3 The location of the eigenvalues of Hermitian, skew-Hermitian, and unitary matrices in the complex plane.

If the definitions of an inner product and a norm are generalized, the concept of orthogonality can be extended to include vectors with complex elements. These generalizations have many applications, but they will only be used here to prove the orthogonality of the rows and columns of unitary matrices.

As the norm of a vector is essentially its *length* and so must be nonnegative, the previous definition of a norm in terms of an inner product must be modified in such a way that the inner product and norm of a complex vector coincide with those for a real vector when purely real vectors are considered. This is achieved by introducing the complex conjugate operation into the definition of an inner product.

Inner product of complex vectors

Let $\mathbf{w} = [w_1, w_2, \ldots, w_n]^T$ and $\mathbf{z} = [z_1, z_2, \ldots, z_n]^T$ be two column vectors with complex elements. Then the **inner product** of the column vectors $\mathbf{w}$ and $\mathbf{z}$, again denoted by $\mathbf{w} \cdot \mathbf{z}$, is defined as $\mathbf{w} \cdot \mathbf{z} = \overline{\mathbf{w}}^T \mathbf{z}$, so that

$$\mathbf{w} \cdot \mathbf{z} = \overline{w}_1 z_1 + \overline{w}_2 z_2 + \cdots + \overline{w}_n z_n. \tag{18}$$

Norm of complex vectors

The **norm** of a vector $\mathbf{z}$, again denoted by $\|\mathbf{z}\|$, is defined as the nonnegative number

$$\begin{aligned} \|\mathbf{z}\| &= (\mathbf{z} \cdot \mathbf{z})^{1/2} = (\overline{\mathbf{z}}^T \mathbf{z})^{1/2} \\ &= (\overline{z}_1 z_1 + \overline{z}_2 z_2 + \cdots + \overline{z}_n z_n)^{1/2} \\ &= (|z_1|^2 + |z_2|^2 + \cdots + |z_n|^2)^{1/2}. \end{aligned} \tag{19}$$

It can be seen from the preceding definition that the inner product of two arbitrary complex vectors is a complex number. However, the definition of the norm of a complex vector $\mathbf{z}$ is a real nonnegative number, as would be expected.

EXAMPLE 4.17 If $\mathbf{w} = [1 + 2i,\ 3 - i,\ i]^T$ and $\mathbf{z} = [2 + i,\ 1 - i,\ 1 + 3i]^T$, find $\mathbf{w} \cdot \mathbf{z}$ and $\|\mathbf{z}\|$.

Solution $\mathbf{w} \cdot \mathbf{z} = (\overline{1 + 2i})(2 + i) + (\overline{3 - i})(1 - i) + \overline{i}(1 + 3i) = 11 - 6i$, and $\|\mathbf{z}\| = [|2 + i|^2 + |1 - i|^2 + |1 + 3i|^2]^{1/2} = 17^{1/2}$. ■

We are now in a position to generalize the concept of an orthonormal system of real vectors to a system of complex vectors that will be called a *unitary* system if the vectors satisfy the following conditions.

A unitary system

A set of complex vectors $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n$ is said to form a **unitary system** if

$$\mathbf{z}_i \cdot \mathbf{z}_j = \overline{\mathbf{z}}_i^T \mathbf{z}_j = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j. \end{cases} \tag{20}$$

THEOREM 4.9

The eigenvectors of a unitary matrix The rows and columns of a unitary matrix each form a unitary system of vectors.

Proof By definition the $n \times n$ matrix $\mathbf{U}$ is unitary if $\overline{\mathbf{U}}^{\mathrm{T}} = \mathbf{U}^{-1}$, so that $\overline{\mathbf{U}}^{\mathrm{T}}\mathbf{U} = \mathbf{I}$. The element in the ith row and jth column of $\mathbf{I}$ is the inner product $\mathbf{x}_i \cdot \mathbf{x}_j = \overline{\mathbf{x}}_i^{\mathrm{T}}\mathbf{x}_j$, where $\mathbf{x}_i$ and $\mathbf{x}_j$ are the ith and jth columns of $\mathbf{U}$. Consequently,

$$\overline{\mathbf{x}}_i^{\mathrm{T}}\mathbf{x}_j = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j, \end{cases}$$

showing that the columns of $\mathbf{U}$ form a unitary system. The rows also form a unitary system, because taking the transpose of $\overline{\mathbf{U}}^{\mathrm{T}}\mathbf{U}$ we find that $(\overline{\mathbf{U}}^{\mathrm{T}}\mathbf{U})^{\mathrm{T}} = \mathbf{U}^{\mathrm{T}}\overline{\mathbf{U}} = \mathbf{I}^{\mathrm{T}} = \mathbf{I}$. ■

Summary

Matrices with complex elements arise in a variety of different applications, and from among these matrices, the most important are Hermitian, skew-Hermitian, and unitary matrices. Hermitian and skew-Hermitian matrices are the complex analogues of real symmetric and skew-symmetric matrices, respectively, and unitary matrices are the complex analogue of real orthogonal matrices. This section derived and illustrated by means of examples the most important properties of these matrices, and then introduced the inner product and norm of matrices with complex elements.

EXERCISES 4.3

In Exercises 1 through 4 write the given matrix as the sum of an Hermitian and a skew-Hermitian matrix.

1. $\begin{bmatrix} 1+i & 3+i & 3+2i \\ -1+3i & 2 & 4+i \\ -3-2i & 2+3i & 4+2i \end{bmatrix}$.

2. $\begin{bmatrix} 0 & 3+i & 1+2i \\ 1-5i & 1+i & 2 \\ 1+4i & -2i & 3 \end{bmatrix}$.

3. $\begin{bmatrix} 4-2i & 1+i & 2+2i \\ -1-3i & 1+2i & 4 \\ 0 & 2 & 0 \end{bmatrix}$.

4. $\begin{bmatrix} 3+i & 4-i & 5+2i \\ 2+i & 1+2i & 2 \\ -1 & 2i & 4-i \end{bmatrix}$.

In Exercises 5 through 8 find the eigenvalues of the Hermitian matrices and hence confirm the result of Theorem 4.8(a) that they are real.

5. $\begin{bmatrix} 1 & 2-i \\ 2+i & 2 \end{bmatrix}$.

6. $\begin{bmatrix} 2 & 2+2i \\ 1-2i & 3 \end{bmatrix}$.

7. $\begin{bmatrix} 3 & 2-3i \\ 2+3i & 1 \end{bmatrix}$.

8. $\begin{bmatrix} -4 & 2-2i \\ 2+2i & 3 \end{bmatrix}$.

In Exercises 9 through 12 find the eigenvalues of the skew-Hermitian matrices and hence confirm the result of Theorem 4.8(b) that they are purely imaginary.

9. $\begin{bmatrix} i & 3+i \\ -3+i & 2i \end{bmatrix}$.

10. $\begin{bmatrix} 3i & 2-i \\ -2-i & 0 \end{bmatrix}$.

11. $\begin{bmatrix} 0 & 3+2i \\ -3+2i & 0 \end{bmatrix}$.

12. $\begin{bmatrix} 4i & 2+3i \\ -2+3i & i \end{bmatrix}$.

13. Show the following matrix is unitary:

$$\begin{bmatrix} 1/\sqrt{2} & -i/\sqrt{2} \\ i/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}.$$

In Exercises 14 and 15 show the matrices are unitary, find their eigenvalues and eigenvectors, and confirm that the eigenvalues all lie on the unit circle.

14. $\begin{bmatrix} (i-1)/\sqrt{2} & (1-i)/\sqrt{2} \\ (i-1)/\sqrt{2} & (i-1)/\sqrt{2} \end{bmatrix}$.

15. $\begin{bmatrix} (1+i)/\sqrt{2} & -(1+i)/\sqrt{2} \\ (1+i)/\sqrt{2} & (1+i)/\sqrt{2} \end{bmatrix}$.

4.4 Quadratic Forms

A homogeneous polynomial $P(\mathbf{x})$ of degree two of the form

$$P(\mathbf{x}) \equiv a_{11}x_1^2 + a_{22}x_2^2 + \cdots + a_{nn}x_n^2 + 2a_{12}x_1x_2 + 2a_{13}x_1x_3 + \cdots + 2a_{n-1,n}x_{n-1}x_n, \tag{21}$$

real quadratic form

in which the coefficients a_{ij} and the variables in $\mathbf{x}(x_1, x_2, \ldots, x_n)$ are real numbers, is called a **real quadratic form** in the variables $x_1, x_2, \ldots, x_n$. The term *homogeneous* of degree two or, more precisely, *algebraically* homogeneous of degree two, means that each term in P is quadratic in the sense that it involves a product of precisely two of the variables $x_1, x_2, \ldots, x_n$. The terms involving the products x_ix_j with $i \neq j$ are called the **mixed product** or **cross-product terms**.

Real quadratic forms

A real quadratic form $P(\mathbf{x})$ is a homogeneous polynomial in the real variables $x_1, x_2, \ldots, x_n$ of the form shown in (21). If $\mathbf{A}$ is a real symmetric $n \times n$ matrix and $\mathbf{x}$ is an n-element column vector defined as

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \quad \text{and} \quad \mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{12} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ a_{1n} & a_{2n} & \cdots & a_{nn} \end{bmatrix}, \tag{22}$$

then $P(\mathbf{x})$ can be written in the matrix form

$$P(\mathbf{x}) \equiv \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}. \tag{23}$$

There is no loss of generality in requiring $\mathbf{A}$ to be a symmetric matrix, because if the coefficient of a cross-product term x_ix_j equals b_{ij}, this can always be rewritten as $b_{ij} = 2a_{ij}$ allowing the terms a_{ij} to be positioned symmetrically about the leading diagonal, as shown in the matrix $\mathbf{A}$ in (22). Exercise 30 at the end of this section shows how the definition of a real quadratic form can be extended to any real $n \times n$ matrix.

EXAMPLE 4.18

Express the quadratic form

$$P(\mathbf{x}) \equiv 3x_1^2 - 2x_2^2 + 4x_3^2 + x_1x_2 + 3x_1x_3 - 2x_2x_3$$

as the matrix product $P(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$.

Solution By defining $\mathbf{x}$ and $\mathbf{A}$ as

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} 3 & 1/2 & 3/2 \\ 1/2 & -2 & -1 \\ 3/2 & -1 & 4 \end{bmatrix},$$

we can write $P(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$. ■

Quadratic forms arise in various ways; for example, in mechanics a quadratic form can describe the ellipsoid of inertia of a solid body, the angular momentum of a solid body rotating about an axis, and the kinetic energy of a system of moving particles. Other areas in which quadratic forms occur include the geometry of conics in two space dimensions and of quadrics in three space dimensions, optimization problems, crystallography, and in the classification of partial differential equations (see Chapter 18).

We now give a general definition of a quadratic form that allows both the matrix $\mathbf{A}$ and the vector $\mathbf{x}$ to contain complex elements.

General quadratic forms

quadratic form and vectors with complex elements

Let the elements of an $n \times n$ matrix $\mathbf{A} = [a_{ij}]$ and an n-element column vector $\mathbf{z}$ be complex numbers. Then a **quadratic form** $P(\mathbf{z})$ involving the variables $z_1, z_2, \ldots, z_n$ of vector $\mathbf{z}$ is an expression of the form

$$P(\mathbf{z}) = \overline{\mathbf{z}}^{\mathrm{T}}\mathbf{A}\mathbf{z} = \sum_{i=1,\,j=1}^{n} a_{ij}\overline{z}_i z_j. \tag{24}$$

This definition is seen to include real quadratic forms, because when the elements of $\mathbf{A}$ and $\mathbf{z}$ are real, result (24) reduces to the real quadratic form defined in (23).

The structure of a quadratic form becomes clearer if a change of variables is made that removes the mixed product terms, leaving only the squared terms. This is called the **reduction** of the quadratic form to its **standard form**, also known as its **canonical form**. The next theorem shows how such a simplification can be achieved.

THEOREM 4.10

how to reduce a quadratic form to a sum of squares

Reduction of a quadratic form Let the $n \times n$ real symmetric matrix $\mathbf{A}$ have the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$, and let $\mathbf{Q}$ be an orthogonal matrix that diagonalizes $\mathbf{A}$, so that $\mathbf{Q}^{\mathrm{T}}\mathbf{A}\mathbf{Q} = \mathbf{D}$, where $\mathbf{D}$ is a diagonal matrix with the eigenvalues of $\mathbf{A}$ as the elements on its leading diagonal. Then the change of variable $\mathbf{x} = \mathbf{Q}\mathbf{y}$, involving the column vectors $\mathbf{x} = [x_1, x_2, \ldots, x_n]^{\mathrm{T}}$ and $\mathbf{y} = [y_1, y_2, \ldots, y_n]^{\mathrm{T}}$, transforms the real quadratic form $P(\mathbf{x}) \equiv \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$ into the **standard form**

$$P(\mathbf{x}) \equiv \sum_{i=1,\,j=1}^{n} a_{ij}x_i x_j = \lambda_1 y_1^2 + \lambda_2 y_2^2 + \cdots + \lambda_n y_n^2.$$

Proof The proof uses the fact that because $\mathbf{Q}$ is an orthogonal matrix, $\mathbf{Q}^{\mathrm{T}}\mathbf{A}\mathbf{Q} = \mathbf{D}$. Substituting $\mathbf{x} = \mathbf{Q}\mathbf{y}$ into the real quadratic form $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$ gives

$$\begin{aligned} P(\mathbf{x}) \equiv \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} &= (\mathbf{Q}\mathbf{y})^{\mathrm{T}}\mathbf{A}\mathbf{Q}\mathbf{y} \\ &= \mathbf{y}^{\mathrm{T}}\mathbf{Q}^{\mathrm{T}}\mathbf{A}\mathbf{Q}\mathbf{y} \\ &= \mathbf{y}^{\mathrm{T}}\mathbf{D}\mathbf{y} = \lambda_1 y_1^2 + \lambda_2 y_2^2 + \cdots + \lambda_n y_n^2. \end{aligned}$$

■

It follows immediately from Theorem 4.10 that the standard form of $P(\mathbf{x})$ is determined once the eigenvalues of $\mathbf{A}$ are known and, when needed, the transformation of coordinates between $\mathbf{x}$ and $\mathbf{y}$ is given by $\mathbf{x} = \mathbf{Q}\mathbf{y}$ or, equivalently, by $\mathbf{y} = \mathbf{Q}^{\mathrm{T}}\mathbf{x}$.

The next example provides a geometrical interpretation of Theorem 4.10 in the context of rigid body mechanics. In order to understand its implications it is necessary to know that if an origin O is taken at an arbitrary point inside a solid body, and an orthogonal set of axes $O\{x_1, x_2, x_3\}$ is located at O, nine moments and products of inertia of the body can be defined relative to these axes and displayed in the form of a 3×3 inertia matrix. The moment of inertia of the body about any line passing through the origin O is proportional to the length of the segment of the line that lies between O and the point where it intersects a three-dimensional surface defined by a quadratic form determined by the inertia matrix.

When the surface determined by the inertia matrix is scaled so the length of the line from O to its point of intersection with the surface equals the reciprocal of the moment of inertia about that line, the surface is called the **ellipsoid of inertia**. If the orientation of the $O\{x_1, x_2, x_3\}$ axes is chosen arbitrarily, the resulting quadratic form will be complicated by the presence of mixed product terms, but a suitable rotation of the axes can always remove these terms and lead to the most convenient orientation of the new system of axes $O\{y_1, y_2, y_3\}$. In the geometry of both conics and quadrics, and also in mechanics, new axes obtained in this way that lead to the elimination of mixed product terms are called the **principal axes**, and it is because of this that Theorem 4.10 is often known as the **principal axes theorem**.

quadratic forms and principal axes

EXAMPLE 4.19

The ellipsoid of inertia of a solid body is given by

$$P(\mathbf{x}) \equiv 4x_1^2 + 4x_2^2 + x_3^2 - 2x_1x_2.$$

Find its standard form in terms of a new orthogonal set of axes $O\{y_1, y_2, y_3\}$, and find the linear transformation that connects the two sets of coordinates.

Solution The quadratic form $P(\mathbf{x})$ can be written as $\mathbf{x}^T\mathbf{A}\mathbf{x}$ by defining

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \quad \text{and} \quad \mathbf{A} = \begin{bmatrix} 4 & -1 & 0 \\ -1 & 4 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The eigenvalues of $\mathbf{A}$ are $\lambda_1 = 1$, $\lambda_2 = 5$, and $\lambda_3 = 3$, so the standard form of $P(\mathbf{x})$ is

$$P(\mathbf{x}) \equiv y_1^2 + 5y_2^2 + 3y_3^2.$$

The eigenvalues and corresponding normalized eigenvectors of $\mathbf{A}$ are

$$\lambda_1 = 1, \quad \hat{\mathbf{x}}_1 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad \lambda_2 = 5, \quad \hat{\mathbf{x}}_2 = \begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \\ 0 \end{bmatrix}, \quad \lambda_3 = 3, \quad \hat{\mathbf{x}}_3 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \\ 0 \end{bmatrix},$$

so the orthogonal diagonalizing matrix for $\mathbf{A}$ is

$$\mathbf{Q} = \begin{bmatrix} 0 & -1/\sqrt{2} & 1/\sqrt{2} \\ 0 & 1/\sqrt{2} & 1/\sqrt{2} \\ 1 & 0 & 0 \end{bmatrix},$$

and the change of variables between $\mathbf{x}$ and $\mathbf{y}$ determined by $\mathbf{x} = \mathbf{Q}\mathbf{y}$ becomes

$$x_1 = (-y_2 + y_3)/\sqrt{2}, \quad x_2 = (y_2 + y_3)/\sqrt{2}, \quad x_3 = y_1.$$

The equation $P(\mathbf{x}) =$ constant is seen to be an *ellipsoid* for which $O\{y_1, y_2, y_3\}$ are the *principal axes*. ■

EXAMPLE 4.20

Reduce the quadratic part of the following expression to its standard form involving the principal axes O$\{y_1, y_2\}$, and hence find the form taken by the complete expression in terms of y_1 and y_2:

$$x_1^2 + 4x_1x_2 + 4x_2^2 + x_1 - 2x_2.$$

Solution The quadratic part of the expression is $x_1^2 + 4x_1x_2 + 4x_2^2$, and this can be expressed in the form $\mathbf{x}^T\mathbf{A}\mathbf{x}$ by setting

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \text{and} \quad \mathbf{A} = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}.$$

The eigenvalues and eigenvectors of $\mathbf{A}$ are

$$\lambda_1 = 5, \quad \mathbf{x}_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad \text{and} \quad \lambda_2 = 0, \quad \mathbf{x}_2 = \begin{bmatrix} -2 \\ 1 \end{bmatrix},$$

so the orthogonal diagonalizing matrix is

$$\mathbf{Q} = \begin{bmatrix} 1/\sqrt{5} & -2/\sqrt{5} \\ 2/\sqrt{5} & 1/\sqrt{5} \end{bmatrix} \quad \text{and} \quad \mathbf{D} = \begin{bmatrix} 5 & 0 \\ 0 & 0 \end{bmatrix}.$$

Making the variable change $\mathbf{x} = \mathbf{Q}\mathbf{y}$ shows the standard form of the quadratic terms to be $5y_1^2$. The variables x_1 and x_2 are related to y_1 and y_2 by the expressions $\mathbf{x}_1 = y_1/\sqrt{5} - 2y_2/\sqrt{5}$ and $x_2 = 2y_1/\sqrt{5} + y_2/\sqrt{5}$, so $x_1 - 2x_2 = -(3y_1 + 4y_2)/\sqrt{5}$. In terms of the principal axes involving the coordinates y_1 and y_2, the complete expression $x_1^2 + 4x_1x_2 + 4x_2^2 + x_1 - 2x_2$ reduces to

$$x_1^2 + 4x_1x_2 + 4x_2^2 + x_1 - 2x_2 = 5y_1^2 - (3y_1 + 4y_2)/\sqrt{5}.$$

■

Quadratic forms $P(\mathbf{x})$ are classified according to the behavior of the sign of $P(\mathbf{x})$ when $\mathbf{x}$ is allowed to take all possible values. In terms of vector spaces, this amounts to saying that if the vector $\mathbf{x}$ in $P(\mathbf{x})$ is an n vector, then $\mathbf{x} \in R^n$.

how to classify quadratic forms

Classification of quadratic forms

Let $P(\mathbf{x})$ be a quadratic form. Then:

1. $P(\mathbf{x})$ is said to be **positive definite** if $P(\mathbf{x}) > 0$ for all $\mathbf{x} \neq \mathbf{0}$ in R^n, with $P(\mathbf{x}) = 0$ if, and only if, $\mathbf{x} = \mathbf{0}$. $P(\mathbf{x})$ is said to be **negative definite** if in this definition the inequality sign $>$ is replaced by $<$.
2. $P(\mathbf{x})$ is said to be **positive semidefinite** if $P(\mathbf{x}) \geq 0$ for all $\mathbf{x} \neq \mathbf{0}$ in R^n, and to be **negative semidefinite** if in this definition the inequality sign $\geq$ is replaced by $\leq$.
3. $P(\mathbf{x})$ is said to be **indefinite** if it satisfies none of the above conditions.

It is an immediate consequence of Theorem 4.10 that if $P(\mathbf{x})$ is associated with a real symmetric matrix $\mathbf{A}$, then:

(a) $P(\mathbf{x})$ is positive definite if all the eigenvalues of $\mathbf{A}$ are positive, and it is negative definite if all the eigenvalues of $\mathbf{A}$ are negative.

(b) $P(\mathbf{x})$ is positive semidefinite if all the eigenvalues of $\mathbf{A}$ are nonnegative, and it is negative semidefinite if all the eigenvalues of $\mathbf{A}$ are nonpositive. So, in each semidefinite case, one or more of the eigenvalues may be zero.

(c) $P(\mathbf{x})$ is indefinite if at least one eigenvalue is opposite in sign to the others. In this case, depending on the choice of $\mathbf{x}$, $P(\mathbf{x})$ may be either positive or negative.

EXAMPLE 4.21 The following are examples of different types of standard forms associated with a 3×3 matrix:

$x_1^2 + 2x_2^2 + 5x_3^2$ is positive definite;
$-(2x_1^2 + 7x_2^2 + 4x_3^2)$ is negative definite;
$4x_1^2 + 3x_3^2$ is positive semidefinite (it is positive, but irrespective of the value of $x_2 \neq 0$ it can vanish when $\mathbf{x} \neq \mathbf{0}$);
$-(2x_1^2 + x_3^2)$ is negative semidefinite (it is negative, but irrespective of the value of $x_2 \neq 0$ it can vanish when $\mathbf{x} \neq \mathbf{0}$);
$3x_1^2 - 2x_2^2 + x_3^2$ is indefinite (it can be positive or negative). ■

Further, and more detailed, information relating to the material in Sections 4.1 to 4.4 is to be found in the appropriate chapters of references [2.1] and [2.5] to [2.12].

Summary

A real quadratic form involving the n real variables $x_1, x_2, \ldots, x_n$ is a homogeneous polynomial of degree two in these variables. Such forms arise in many different ways, one of which occurs in optimization problems where a reduction to a sum of squares simplifies the task of finding an optimum least squares solution. In this section it was shown that a real quadratic form arises when studying the mechanics of solid bodies, since there a set of principal axes $O\{x_1, x_2, x_3\}$ is used to simplify the description of the body in terms of its inertia about each of the three axes. The reduction of a quadratic form to a sum of squares both simplifies the analysis of its properties and also enables it to be classified as being positive or negative definite, semipositive or seminegative, or of indefinite type, all of which classifications have important implications in applications.

EXERCISES 4.4

In Exercises 1 through 6 find the symmetric matrix $\mathbf{A}$ that is associated with the given quadratic form.

1. $x_1^2 + 4x_1x_3 - 6x_2x_3 + 3x_2^2 - 2x_3^2$.
2. $5x_1^2 - 2x_2^2 - 5x_3^2 - 4x_2x_3$.
3. $-2x_1^2 + 3x_2^2 - 2x_1x_3 + 4x_2x_3$.
4. $x_1^2 + 3x_2^2 - 2x_1x_2 + 4x_2x_4 - 2x_3x_4 + x_3^2 + 6x_4^2$.
5. $3x_1^2 - 4x_1x_2 - 6x_2x_3 - 2x_2x_4 + 2x_3^2 + 8x_4^2$.
6. $x_1^2 + x_2^2 + 4x_3^2 - 3x_4^2 - x_1x_2 + 2x_2x_4 + 2x_3x_4$.

In Exercises 7 through 10 write down the quadratic form associated with the given matrix.

7. $\begin{bmatrix} 2 & 4 & 4 & 0 \\ 4 & 1 & 2 & 1 \\ 4 & 2 & -1 & 2 \\ 0 & 1 & 2 & 3 \end{bmatrix}$.

8. $\begin{bmatrix} 1 & -3 & 2 & 1 \\ -3 & 2 & 0 & 2 \\ 2 & 0 & -3 & 0 \\ 1 & 2 & 0 & 4 \end{bmatrix}$.

9. $\begin{bmatrix} 0 & 2 & -4 & 2 \\ 2 & 3 & 1 & 0 \\ -4 & 1 & 2 & 1 \\ 2 & 0 & 1 & 7 \end{bmatrix}$.

10. $\begin{bmatrix} 1 & -2 & 4 & 3 \\ -2 & 3 & 1 & 2 \\ 4 & 1 & 5 & 0 \\ 3 & 2 & 0 & 3 \end{bmatrix}$.

In Exercises 11 through 18 use hand computation to reduce the quadratic form to its standard form, and use the reduction to classify it. Confirm the reduction by using computer algebra.

11. $(5/2)x_1^2 + x_1x_3 + x_2^2 + (5/2)x_3^2$.
12. $4x_1^2 + x_2^2 + 2x_2x_3 + x_3^2$.
13. $4x_1^2 + 4x_2^2 + 2x_2x_3 + 4x_3^2$.
14. $(3/2)x_1^2 - x_1x_3 + x_2^2 + (3/2)x_3^2$.
15. $(3/2)x_1^2 + x_1x_3 - x_2^2 + (3/2)x_3^2$.
16. $(1/2)x_1^2 + x_1x_3 + 2x_2^2 + (1/2)x_3^2$.
17. $2x_1^2 + x_2^2 - 4x_2x_3 + x_3^2$.
18. $2x_1^2 + 2x_2^2 + 2x_2x_3 + 2x_3^2$.

In Exercises 19 through 24 use computer algebra to reduce the quadratic form on the left to its standard form. Use the result to identify the conic section described by the equation as a circle, an ellipse, or a hyperbola.

19. $3x_1^2 - 6x_1x_2 + 9x_2^2 = 3$.
20. $8x_2^2 - x_1^2 + 20x_1x_2 = 12$.

21. $5x_1^2 + 4x_1x_2 - 10x_2^2 = 1.$
22. $10x_1^2 + 2x_1x_2 + 5x_2^2 = 4.$
23. $13x_1^2 + 18x_1x_2 + 10x_2^2 = 9.$
24. $2x_1^2 + 16x_1x_2 + 5x_2^2 = 4.$

In Exercises 25 through 29 use hand computation to reduce the quadratic part of the expression to its standard form involving the principal axes $O\{y_1, y_2\}$, and find the form taken by the complete expression in terms of y_1 and y_2. Confirm the reduction by using computer algebra.

25. $x_1^2 + 8x_1x_2 + x_2^2 + 3x_1 - 2x_2.$
26. $x_1^2 - 8x_1x_2 + x_2^2 + 2x_1 + 3x_2.$
27. $-2x_1^2 + 4x_1x_2 + x_2^2 + 4x_1 - x_2.$
28. $(8/5)x_1^2 - (8/5)x_1x_2 + (2/5)x_2^2 + 2x_1 + 4x_2.$
29. $(35/17)x_1^2 + (8/17)x_1x_2 + (50/17)x_2^2 + 4x_2.$
30. By using the definitions of a symmetric and a skew-symmetric matrix, generalize the definition of a quadratic form by proving that the quadratic form associated with any real $n \times n$ matrix $\mathbf{A}$ can be written $\mathbf{x}^T\mathbf{B}\mathbf{x}$, where $\mathbf{B}$ is the symmetric part of $\mathbf{A}$.

4.5 The Matrix Exponential

It is shown in Chapter 6 that the matrix exponential can be used when solving systems of linear first order differential equations. As this approach uses matrix diagonalization when determining what is called the *matrix exponential* involving an arbitrary $n \times n$ diagonalizable matrix, it is convenient to introduce the matrix exponential in this chapter.

To motivate what is to follow, we notice that the first order homogeneous linear differential equation

$$dx/dt = ax \quad (a = \text{constant}) \tag{25}$$

has the general solution

$$x = ce^{at} \tag{26}$$

where c is an arbitrary constant.

Let us now consider the system of n linear first order homogeneous differential equations

$$\begin{aligned} dx_1/dt &= a_{11}x_1 + a_{12}x_2 + \cdots a_{1n}x_n \\ dx_2/dt &= a_{21}x_1 + a_{22}x_2 + \cdots a_{2n}x_n \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ dx_n/dt &= a_{n1}x_1 + a_{n2}x_2 + \cdots a_{nn}x_n \end{aligned} \tag{27}$$

Setting

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \quad \text{and} \quad \mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

allows the system of differential equations in (27) to be written in the matrix form

$$d\mathbf{x}/dt = \mathbf{A}\mathbf{x}, \tag{28}$$

where $d\mathbf{x}/dt = [dx_1/dt, dx_2/dt, \ldots, dx_n/dt]^T$ (see Section 3.2(d)).

As the single differential equation (25) has the solution (26), it is reasonable to ask whether it is possible to express the solution of the system of differential equations in (28) in the form

$$\mathbf{x} = e^{\mathbf{A}t}\mathbf{C}. \tag{29}$$

the matrix exponential

For this to be possible it is necessary to give meaning to the expression $e^{\mathbf{A}t}$, which is called the **matrix exponential**, with t as a parameter. Our objective in the remainder of this section will be to give a brief introduction to the matrix exponential and to use the definition to determine its most important properties in preparation for their use in Chapter 6.

The starting point for this generalization of the exponential function is the familiar result

$$\begin{aligned} e^{at} &= \sum_{m=0}^{\infty} \frac{a^m t^m}{m!} \\ &= 1 + at + \frac{a^2t^2}{2!} + \frac{a^3t^3}{3!} + \cdots. \end{aligned} \tag{30}$$

If $\mathbf{A}$ is an $n \times n$ constant matrix with real coefficients we take as an intuitive definition of the matrix exponential e^{At} the infinite series of matrices

$$e^{\mathbf{A}t} = \mathbf{I} + \mathbf{A}t + \mathbf{A}^2\frac{t^2}{2!} + \mathbf{A}^3\frac{t^3}{3!} + \cdots. \tag{31}$$

In adopting (31) as a possible definition of the matrix exponential, we have set $\mathbf{A}^0 = \mathbf{I}$ and chosen to vary the convention that a scalar multiplier of a matrix is placed in front of the matrix by writing $\mathbf{A}t, \mathbf{A}^2t^2, \ldots,$ instead of $t\mathbf{A}, t^2\mathbf{A}^2, \ldots$. This notation has been adopted to make the appearance of the arguments that follow parallel as closely as possible those for the familiar single real variable case. Some books adopt this convention but make no mention of it, while others adhere strictly to the convention that a scalar multiplier is placed before a matrix and write

$$e^{t\mathbf{A}} = \mathbf{I} + t\mathbf{A} + \frac{t^2}{2!}\mathbf{A}^2 + \frac{t^3}{3!}\mathbf{A}^3 + \cdots.$$

The matrix exponential in (31) is an $n \times n$ matrix, each element of which is an ordinary infinite series. So to show that $e^{\mathbf{A}t}$ is convergent, it will be sufficient to show that an infinite sum of the required form containing the term of greatest absolute value in $\mathbf{A}$ is convergent. Let us consider the matrix product $\mathbf{A}^2$. Then the term $c_{rs}^{(2)}$ in the rth row and sth column of $\mathbf{A}^2$ is $c_{rs}^{(2)} = a_{r1}a_{1s} + a_{r2}a_{2s} + \cdots + a_{rn}a_{ns}$, so if the magnitude of the largest term in $\mathbf{A}$ is M, it follows that $|a_{rs}| \leq M$, and $|c_{rs}^{(2)}| \leq nM^2$. A similar argument shows that if $|c_{rs}^{(3)}|$ is the corresponding term in the matrix $\mathbf{A}^3$, then $c_{rs}^{(3)} = c_{r1}^{(2)}a_{1s} + c_{r2}^{(2)}a_{2s} + \cdots + c_{rn}^{(2)}a_{ns}$ and so $|c_{rs}^{(3)}| \leq n^2M^3$. Either by induction or by inspection, we see that the magnitude of the term $c_{rs}^{(m)}$ in the rth row and sth column of $\mathbf{A}^m$ obeys the inequality $|c_{rs}^{(m)}| \leq n^{m-1}M^m$.

An overestimate of the magnitude of the term in the rth row and sth column of $e^{\mathbf{A}t}$ is provided by the series

$$1 + tM + t^2nM^2/2! + t^3n^2M^3/3! + \cdots + t^mn^{m-1}M^m/m! + \cdots.$$

Setting $u_m = t^m n^{m-1} M^m / m!$ and applying the ratio test shows that for all fixed t

$$L = \lim_{m\to\infty} |u_{m+1}/u_m| = \lim_{m\to\infty} tnM/(m+1) = 0,$$

so the series is absolutely convergent for all fixed t. Thus, (26) serves as a satisfactory definition of the matrix exponential, and because it is absolutely convergent for all fixed t the series can be differentiated and integrated term by term with respect to t.

The matrix exponential

the formal definition of e^{At} and its properties

If $\mathbf{A}$ is an $n \times n$ constant matrix with real coefficients, the **matrix exponential** $e^{\mathbf{A}t}$ is defined by the infinite series

$$e^{\mathbf{A}t} = \mathbf{I}_n + \mathbf{A}t + \mathbf{A}^2\frac{t^2}{2!} + \mathbf{A}^3\frac{t^3}{3!} + \cdots, \tag{32}$$

which is absolutely convergent for all fixed t.

The absolute convergence of the infinite series defining the matrix exponential allows it to be differentiated term by term, so

$$\begin{aligned} d[e^{\mathbf{A}t}]/dt &= \mathbf{A} + \mathbf{A}^2 t + \mathbf{A}^3\frac{t^2}{2!} + \cdots = \mathbf{A}\left\{\mathbf{I} + \mathbf{A}t + \mathbf{A}^2\frac{t^2}{2!} + \mathbf{A}^3\frac{t^3}{3!} + \cdots\right\} \\ &= \mathbf{A}e^{\mathbf{A}t}. \end{aligned}$$

We have established the fundamental result that

$$d[e^{\mathbf{A}t}]/dt = \mathbf{A}e^{\mathbf{A}t}, \tag{33}$$

and hence by repeated differentiation that

$$d^n[e^{\mathbf{A}t}]/dt^n = \mathbf{A}^n e^{\mathbf{A}t}. \tag{34}$$

Setting $t = 1$ in (33) shows that

$$e^{\mathbf{A}} = \mathbf{I} + \mathbf{A} + \mathbf{A}^2\frac{1}{2!} + \mathbf{A}^3\frac{1}{3!} + \cdots, \tag{35}$$

whereas setting $t = 0$ shows that $e^{\mathbf{0}} = \mathbf{I}$.

EXAMPLE 4.22

Find $e^{\mathbf{A}t}$ given that

$$\mathbf{A} = \begin{bmatrix} 3 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & 4 \end{bmatrix}.$$

Solution As $\mathbf{A}$ is a diagonal matrix

$$\mathbf{A}^m = \begin{bmatrix} 3^m & 0 & 0 \\ 0 & (-2)^m & 0 \\ 0 & 0 & 4^m \end{bmatrix},$$

so substituting into (32) gives

$$e^{\mathbf{A}t} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} + \begin{bmatrix} 3 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & 4 \end{bmatrix} t + \begin{bmatrix} 3^2 & 0 & 0 \\ 0 & (-2)^2 & 0 \\ 0 & 0 & 4^2 \end{bmatrix} \frac{t^2}{2!} + \cdots,$$

showing that

$$e^{\mathbf{A}t} = \begin{bmatrix} \displaystyle\sum_{m=0}^{\infty} \frac{3^m t^m}{m!} & 0 & 0 \\ 0 & \displaystyle\sum_{m=0}^{\infty} \frac{(-2)^m t^m}{m!} & 0 \\ 0 & 0 & \displaystyle\sum_{m=0}^{\infty} \frac{4^m t^m}{m!} \end{bmatrix} = \begin{bmatrix} e^{3t} & 0 & 0 \\ 0 & e^{-2t} & 0 \\ 0 & 0 & e^{4t} \end{bmatrix}.$$

EXAMPLE 4.23 Find $e^{\mathbf{A}}$ and $e^{\mathbf{A}t}$, and show by direct differentiation that $d[e^{\mathbf{A}t}]/dt = \mathbf{A}e^{\mathbf{A}t}$, given that

$$\mathbf{A} = \begin{bmatrix} 0 & 2 & 1 & 1 \\ 0 & 0 & 3 & -2 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Solution

$$\mathbf{A}^2 = \begin{bmatrix} 0 & 0 & 6 & -3 \\ 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{A}^3 = \begin{bmatrix} 0 & 0 & 0 & 6 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad \text{and} \quad \mathbf{A}^n = \mathbf{0} \text{ for } n > 3.$$

Substituting into (32) and adding the scaled matrices gives

$$e^{\mathbf{A}t} = \begin{bmatrix} 1 & 2t & t + 3t^2 & t - (3/2)t^2 + t^3 \\ 0 & 1 & 3t & -2t + (3/2)t^2 \\ 0 & 0 & 1 & t \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Setting $t = 1$ in this result, we find that

$$e^{\mathbf{A}} = \begin{bmatrix} 1 & 2 & 4 & 1/2 \\ 0 & 1 & 3 & -1/2 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Differentiation of the terms in the matrix $e^{\mathbf{A}t}$ gives

$$d[e^{\mathbf{A}t}]/dt = \begin{bmatrix} 0 & 2 & 1 + 6t & 1 - 3t + 3t^2 \\ 0 & 0 & 3 & -2 + 3t \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

and as this is equal to $Ae^{\mathbf{A}t}$, it confirms the result $d[e^{\mathbf{A}t}]/dt = \mathbf{A}e^{\mathbf{A}t}$.

It was possible to sum the infinite series of matrices in Example 4.22 because only a diagonal matrix was involved, so its powers could be determined immediately. The situation was different in Example 4.23 because $\mathbf{A}^n = \mathbf{0}$ for $n > 3$ so that only a finite sum of matrices was involved. Matrices such as those in Example 4.23, which vanish when raised to a finite power, are called **nilpotent** matrices.

If $\mathbf{A}$ is neither diagonal nor nilpotent, but is diagonalizable, in order to determine $\mathbf{A}^m$ it is first necessary to find the diagonalizing matrix $\mathbf{P}$ for $\mathbf{A}$. Then, if $\mathbf{D}$ is the diagonalized form of $\mathbf{A}$, so that $\mathbf{D} = \mathbf{P}^{-1}\mathbf{AP}$, it follows that $\mathbf{A} = \mathbf{PDP}^{-1}$ and

$$\mathbf{A}^2 = (\mathbf{PDP}^{-1})(\mathbf{PDP}^{-1}) = \mathbf{PD}^2\mathbf{P}^{-1}, \quad \mathbf{A}^3 = \mathbf{AA}^2 = (\mathbf{PDP}^{-1})(\mathbf{PD}^2\mathbf{P}^{-1}) = \mathbf{PD}^3\mathbf{P}^{-1},$$

so that in general,

$$\mathbf{A}^m = \mathbf{PD}^m\mathbf{P}^{-1}.$$

Using this result in the matrix exponential gives

$$e^{\mathbf{A}t} = \mathbf{I} + (\mathbf{PDP}^{-1})t + \mathbf{PD}^2\mathbf{P}^{-1}\frac{t^2}{2!} + \cdots,$$

and writing $\mathbf{I} = \mathbf{PP}^{-1}$ reduces this to

$$e^{\mathbf{A}t} = \mathbf{P}\left\{\mathbf{I}_n + \mathbf{D}t + \mathbf{D}^2\frac{t^2}{2!} + \mathbf{D}^3\frac{t^3}{3!} + \cdots\right\}\mathbf{P}^{-1}. \tag{36}$$

The form of $e^{\mathbf{A}}$ follows directly from this by setting $t = 1$.

EXAMPLE 4.24 Determine $e^{\mathbf{A}t}$ given that

$$\mathbf{A} = \begin{bmatrix} -2 & -3 \\ 6 & 7 \end{bmatrix},$$

and use the result to find $e^{\mathbf{A}}$.

Solution The eigenvalues and eigenvectors of $\mathbf{A}$ are

$$\lambda_1 = 1, \quad \mathbf{x}_1 = \begin{bmatrix} -1 \\ 1 \end{bmatrix} \quad \text{and} \quad \lambda_2 = 4, \quad \mathbf{x}_2 = \begin{bmatrix} 1 \\ -2 \end{bmatrix},$$

so the diagonalizing matrix

$$\mathbf{P} = \begin{bmatrix} -1 & 1 \\ 1 & -2 \end{bmatrix} \quad \text{and} \quad \mathbf{P}^{-1} = \begin{bmatrix} -2 & -1 \\ -1 & -1 \end{bmatrix}, \quad \text{while } \mathbf{D} = \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix}.$$

Substituting these matrices into (36) gives

$$e^{\mathbf{A}t} = \mathbf{P}\left[\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix}t + \begin{bmatrix} 1 & 0 \\ 0 & 4^2 \end{bmatrix}\frac{t^2}{2!} + \begin{bmatrix} 1 & 0 \\ 0 & 4^3 \end{bmatrix}\frac{t^3}{3!} + \cdots\right]\mathbf{P}^{-1}$$

$$= \mathbf{P}\begin{bmatrix} e^t & 0 \\ 0 & e^{4t} \end{bmatrix}\mathbf{P}^{-1} = \begin{bmatrix} (2e^t - e^{4t}) & (e^t - e^{4t}) \\ (2e^{4t} - 2e^t) & (2e^{4t} - e^t) \end{bmatrix}.$$

Finally, setting $t = 1$ we find that

$$e^{\mathbf{A}} = \begin{bmatrix} (2e - e^4) & (e - e^4) \\ (2e^4 - 2e) & (2e^4 - e) \end{bmatrix}.$$

So far, the properties of the matrix exponential have closely paralleled those of the ordinary exponential, but there are significant differences, one of the most important being that in general, even when $\mathbf{A} + \mathbf{B}$ is defined, $e^{\mathbf{A}}e^{\mathbf{B}} \neq e^{(\mathbf{A}+\mathbf{B})}$. To determine under what conditions the equality is true, we consider the matrix exponentials $e^{\mathbf{A}t}e^{\mathbf{B}t}$ and $e^{(\mathbf{A}+\mathbf{B})t}$ and require their derivatives to be equal when $t = 0$.

Differentiating each expression once with respect to t gives

$$d[e^{\mathbf{A}t}e^{\mathbf{B}t}]/dt = \mathbf{A}e^{\mathbf{A}t}e^{\mathbf{B}t} + e^{\mathbf{A}t}\mathbf{B}e^{\mathbf{B}t} \quad \text{and} \quad d[e^{(\mathbf{A}+\mathbf{B}t)}]/dt = (\mathbf{A} + \mathbf{B})e^{(\mathbf{A}+\mathbf{B})t},$$

and these are seen to be equal when $t = 0$. Next, computing $d^2[e^{\mathbf{A}t}e^{\mathbf{B}t}]/dt^2$ and $d^2[e^{(\mathbf{A}+\mathbf{B})t}]/dt^2$, we obtain

$$d^2[e^{\mathbf{A}t}e^{\mathbf{B}t}]/dt^2 = \mathbf{A}^2e^{\mathbf{A}t}e^{\mathbf{B}t} + 2\mathbf{A}e^{\mathbf{A}t}\mathbf{B}e^{\mathbf{B}t} + e^{\mathbf{A}t}\mathbf{B}^2e^{\mathbf{B}t}$$

and

$$d^2[e^{(\mathbf{A}+\mathbf{B})t}]/dt^2 = (\mathbf{A} + \mathbf{B})^2e^{(\mathbf{A}+\mathbf{B})t} = (\mathbf{A}^2 + \mathbf{AB} + \mathbf{BA} + \mathbf{B}^2)e^{(\mathbf{A}+\mathbf{B})t}.$$

Setting $t = 0$ shows that these two expressions are only equal if $\mathbf{AB} = \mathbf{BA}$; that is, the matrices $\mathbf{A}$ and $\mathbf{B}$ must *commute*, and the same condition applies when all higher order derivatives are considered. This has established the fundamental result that

when does $e^{\mathbf{A}}e^{\mathbf{B}} = e^{(\mathbf{A}+\mathbf{B})}$

$$e^{\mathbf{A}}e^{\mathbf{B}} = e^{(\mathbf{A}+\mathbf{B})} \quad \text{if, and only if, } \mathbf{AB} = \mathbf{BA}. \tag{37}$$

Replacing $\mathbf{B}$ by $-\mathbf{A}$ in (37) gives

$$e^{\mathbf{A}}e^{-\mathbf{A}} = e^{\mathbf{0}} = \mathbf{I}, \tag{38}$$

from which we see, as would be expected, that $e^{-\mathbf{A}}$ is the inverse of $e^{\mathbf{A}}$, and also that as $e^{-\mathbf{A}}$ is nonsingular it always exists. This parallels the real variable situation, because e^{-x} exists for all finite x.

Having arrived at a satisfactory definition of $e^{\mathbf{A}t}$ and determined its derivatives, we are now in a position to define the **antiderivative** $\int e^{\mathbf{A}t}\,dt$ as the matrix obtained by integrating each element of $e^{\mathbf{A}t}$ with respect to t, it being understood that when this is done an arbitrary constant $n \times n$ matrix must always be added to the result representing the arbitrary additative constant of integration that arises when each term of $e^{\mathbf{A}t}$ is integrated.

EXAMPLE 4.25 Find $\int e^{\mathbf{A}t}\,dt$ given that $\mathbf{A}$ is the matrix in Example 4.21.

Solution It was shown in Example 4.21 that if

$$\mathbf{A} = \begin{bmatrix} 3 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & 4 \end{bmatrix} \quad \text{then } e^{\mathbf{A}t} = \begin{bmatrix} e^{3t} & 0 & 0 \\ 0 & e^{-2t} & 0 \\ 0 & 0 & e^{4t} \end{bmatrix},$$

so that

$$\int e^{\mathbf{A}t}\,dt = \begin{bmatrix} e^{3t}/3 + c_1 & 0 & 0 \\ 0 & -e^{-2t}/2 + c_2 & 0 \\ 0 & 0 & e^{4t}/4 + c_3 \end{bmatrix}$$
$$= \begin{bmatrix} e^{3t}/3 & 0 & 0 \\ 0 & -e^{-2t}/2 & 0 \\ 0 & 0 & e^{4t}/4 \end{bmatrix} + \begin{bmatrix} c_1 & 0 & 0 \\ 0 & c_2 & 0 \\ 0 & 0 & c_3 \end{bmatrix},$$

where c_1, c_2, and c_3 are arbitrary constants. ■

Applications of the matrix exponential to ordinary differential equations are to be found in reference [3.15].

Summary

The matrix exponential $e^{\mathbf{A}t}$ arises as the natural extension of the exponential function when solving a system of linear first order constant coefficient differential equations in the matrix form $d\mathbf{x}/dt = \mathbf{A}\mathbf{x}$. This section has described how $e^{\mathbf{A}t}$ can be calculated in simple cases and shown that $e^{\mathbf{A}}e^{\mathbf{B}} = e^{\mathbf{A}+\mathbf{B}}$ if, and only if, $\mathbf{AB} = \mathbf{BA}$. A different way of finding $e^{\mathbf{A}t}$ using the Laplace transform is given later in Section 7.3(b).

EXERCISES 4.5

1. Given that

$$\mathbf{A} = \begin{bmatrix} 0 & 3 & 1 & 0 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

show that it is nilpotent and find the smallest power for which $\mathbf{A}^n = \mathbf{0}$.

2. Given that

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 2 & 2 \\ 0 & 0 & 3 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

find $e^{\mathbf{A}t}$.

3. Given that

$$\mathbf{A} = \begin{bmatrix} 0 & 2 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} 0 & 0 \\ 3 & 0 \end{bmatrix},$$

show that $\mathbf{A}$ and $\mathbf{B}$ do not commute, and by finding $e^{\mathbf{A}t}$, $e^{\mathbf{B}t}$, and $e^{(\mathbf{A}+\mathbf{B})t}$, verify that $e^{\mathbf{A}t}e^{\mathbf{B}t} \neq e^{(\mathbf{A}+\mathbf{B})t}$.

In Exercises 4 through 9, find $e^{\mathbf{A}t}$.

4. $\mathbf{A} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$.

5. $\mathbf{A} = \begin{bmatrix} m & 0 \\ 0 & n \end{bmatrix}$.

6. $\mathbf{A} = \begin{bmatrix} 0 & -c \\ c & 0 \end{bmatrix}$.

7. $\mathbf{A} = \begin{bmatrix} -2 & 2 \\ 2 & 1 \end{bmatrix}$.

8. $\mathbf{A} = \begin{bmatrix} 3 & -2 & 2 \\ 6 & -4 & 6 \\ 2 & -1 & 3 \end{bmatrix}$.

9. $\mathbf{A} = \begin{bmatrix} 0 & 1 & -2 \\ 2 & -1 & 2 \\ 2 & -2 & 4 \end{bmatrix}$.

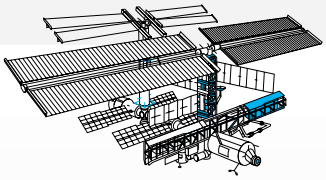
10. By considering the definition of $e^{\mathbf{A}t}$ show, provided the square matrices $\mathbf{A}$ and $\mathbf{B}$ commute, that

$$\mathbf{A}e^{\mathbf{B}t} = e^{\mathbf{B}t}\mathbf{A}.$$

11. By considering the definition of $e^{\mathbf{A}t}$ show that $\int e^{-\mathbf{A}t}\,dt = -\mathbf{A}^{-1}e^{-\mathbf{A}t} + \mathbf{C} = e^{-\mathbf{A}t}\mathbf{A}^{-1} + \mathbf{C}$, where $\mathbf{C}$ is an arbitrary constant matrix that is conformable for addition with $\mathbf{A}$.

12. Show that if the square matrices $\mathbf{A}$ and $\mathbf{B}$ commute, then the binomial theorem takes the form

$$(\mathbf{A}+\mathbf{B})^n = \sum_{k=0}^{n} \binom{n}{k} \mathbf{A}^k \mathbf{B}^{n-k}.$$



CHAPTER 4
TECHNOLOGY PROJECTS

Project 1

Verifying and Using the Cayley–Hamilton Theorem

The purpose of this project is to verify the Cayley–Hamilton theorem in a particular case by constructing an arbitrary 6×6 non-singular matrix $\mathbf{A}$ and, after finding its characteristic polynomial, showing by direct calculation that $\mathbf{A}$ satisfies its own characteristic matrix polynomial equation. The matrix polynomial equation is then to be used to compute the inverse matrix $\mathbf{A}^{-1}$, after which the inverse is to be checked by showing that the product $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$. The project then explores the way in which this approach fails when $\mathbf{A}$ is singular.

1. Construct an arbitrary 6×6 matrix $\mathbf{A}$ and check that $\det \mathbf{A} \neq 0$ to ensure that it has an inverse $\mathbf{A}^{-1}$.
2. Find the characteristic polynomial for matrix $\mathbf{A}$.
3. Show by direct calculation that $\mathbf{A}$ satisfies its own characteristic matrix polynomial equation.
4. Use the characteristic matrix polynomial equation to find $\mathbf{A}^{-1}$, and check its correctness by showing that the product $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$.
5. Replace the last row of $\mathbf{A}$ by the entries in the row above to form a matrix $\mathbf{B}$ that is singular, and find the characteristic polynomial for $\mathbf{B}$.
6. Try to use the characteristic matrix polynomial equation for $\mathbf{B}$ to find $\mathbf{B}^{-1}$, and comment on the way in which this approach fails.

Project 2

Diagonalization of a Matrix

This project involves the diagonalization of a 5×5 matrix $\mathbf{A}$ when two of its five eigenvalues are equal, but there are five linearly independent eigenvectors.

1. Find a diagonalizing matrix for

$$\mathbf{A} = \begin{bmatrix} 13 & 31 & 30 & 51 & -40 \\ 32 & 62 & 64 & 104 & -88 \\ -28 & -56 & -58 & -88 & 80 \\ -17 & -33 & -34 & -55 & 48 \\ -13 & -25 & -26 & -37 & 38 \end{bmatrix}.$$

2. Diagonalize the matrix $\mathbf{B} = \frac{1}{2}\mathbf{A}$, and comment on the relationship between the diagonalizing matrices for $\mathbf{A}$ and $\mathbf{B}$.

Project 3

Orthogonal Vectors Computed by the Gram–Schmidt Method

The purpose of this project is to develop a computer algebra procedure that generalizes the **Gram–Schmidt** process to n-dimensional vectors. The extension is almost immediate and follows from the fact that in the case of three-dimensional vectors one of them, say $\mathbf{a}_1$, was taken as the first vector $\mathbf{u}_1$ of an orthogonal basis, the second vector $\mathbf{u}_2$ was derived from $\mathbf{a}_2$ by subtracting from it the projection of $\mathbf{u}_1$ onto $\mathbf{a}_2$, and, finally, the third vector $\mathbf{u}_3$ was obtained from $\mathbf{a}_3$ by subtracting from it both the projection of $\mathbf{u}_1$ onto $\mathbf{a}_3$ and the projection of $\mathbf{u}_2$ onto $\mathbf{a}_3$.

Starting with a set of n linearly independent vectors $\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n\}$, an orthogonal basis $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ for this space is obtained by extending the preceding method by setting

$$\begin{aligned}
\mathbf{u}_1 &= \mathbf{a}_1 \\
\mathbf{u}_2 &= \mathbf{a}_2 - \frac{\mathbf{a}_2.\mathbf{u}_1}{\mathbf{u}_1.\mathbf{u}_1}\mathbf{u}_1 \\
\mathbf{u}_3 &= \mathbf{a}_3 - \frac{\mathbf{a}_3.\mathbf{u}_1}{\mathbf{u}_1.\mathbf{u}_1}\mathbf{u}_1 - \frac{\mathbf{a}_3.\mathbf{u}_2}{\mathbf{u}_2.\mathbf{u}_2}\mathbf{u}_2 \\
&\vdots \\
\mathbf{u}_n &= \mathbf{a}_n - \frac{\mathbf{a}_n.\mathbf{u}_1}{\mathbf{u}_1.\mathbf{u}_1}\mathbf{u}_1 - \frac{\mathbf{a}_n.\mathbf{u}_2}{\mathbf{u}_2.\mathbf{u}_2}\mathbf{u}_2 - \cdots - \frac{\mathbf{a}_n.\mathbf{u}_{n-1}}{\mathbf{u}_{n-1}.\mathbf{u}_{n-1}}\mathbf{u}_{n-1}.
\end{aligned}$$

Write a computer algebra procedure that reproduces these results step by step for four-dimensional vectors.

Check the procedure by applying it to the set of linearly independent vectors $\mathbf{a}_1 = [-1, -1, 1, 2]$, $\mathbf{a}_2 = [1, 0, 1, -2]$, $\mathbf{a}_3 = [0, 1, -1, -1]$, and $\mathbf{a}_4 = [2, -1, 1, 1]$, and showing that the corresponding set of orthogonal basis vectors is $\mathbf{u}_1 = [-1, -1, 1, 2]$, $\mathbf{u}_2 = [\frac{3}{7}, -\frac{4}{7}, \frac{11}{7}, -\frac{6}{7}]$, $\mathbf{u}_3 = [-\frac{11}{26}, \frac{3}{13}, \frac{3}{26}, -\frac{2}{13}]$, and $\mathbf{u}_4 = [\frac{2}{7}, \frac{4}{7}, \frac{2}{7}, \frac{2}{7}]$.

Define two other sets of linearly independent vectors and, after applying your procedure, verify that the resulting sets of vectors $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \mathbf{u}_4\}$ are orthogonal.

Project 4

Reduction of a Quadratic Form to Standard Form

The purpose of this project is to find a transformation that reduces a given quadratic form in four variables to a sum of squares.

1. Given the quadratic form $x_2^2 - x_1^2 - 2x_1x_2 - 2x_1x_3 + 2x_1x_4 - 2x_3x_4$, find a transformation that reduces it to a sum of squares.
2. Find the simplified quadratic form produced by the transformation in Step 1.

Project 5

The Hubble Space Telescope and Quadratic Forms

When the Hubble space telescope in orbit around the earth is required to photograph a particular nebula it has to be rotated until it is pointing in the correct direction. As it is a rigid body, the kinetic energy W required to rotate it at an angular velocity ω about a suitable axis is given by $W = \frac{1}{2}I\omega^2$, where I is the moment of inertia of the telescope about the axis of rotation. Because the telescope has an irregular shape, the moment of inertia I will depend on the axis of rotation, and a convenient way of representing the value of I about all possible axes through a given point in the telescope is by means of what is called the *ellipsoid of inertia*.

The ellipsoid of inertia for a given rigid body of mass m relative to a fixed point in the body is a three-dimensional plot of the moment of inertia relative to all possible axes of rotation passing through the point. It is shown in texts on mechanics that this plot is an ellipsoidal surface, with the property that the length of the straight line drawn from the center of the ellipsoid to its surface is inversely proportional to the radius of gyration k of the body about that line, where $I = mk^2$.

Given that an ellipsoid of inertia has the form

$$16x^2 - 4xy + 37y^2 - 12xz + 18yz + 11z^2 = 12,$$

use matrix methods to find a linear transformation from the variables x, y, and z to new variables X, Y, and Z that reduces the expression to one of the form

$$\frac{X^2}{a^2} + \frac{Y^2}{b^2} + \frac{Z^2}{c^2} = 1.$$

Hence find the radii of gyration $1/a, 1/b$, and $1/c$ about the principal axes of the ellipsoid that form its three mutually orthogonal axes about which there is symmetry.

Project 6

Dynamical Systems and Logging Operations

Discrete dynamical systems are used to model situations in engineering, control theory, physics, ecology, and elsewhere that can be considered to evolve stage by stage, with each stage dependent on the previous one. For example, a logging operation to supply a saw mill in a specific area of forest, with tree replanting and the availability of a limited supply of logs from outside the area, can be described by a simple dynamical system that models the way the output of cut timber is influenced by the competition between the felling of trees, the importing of a limited amount of logs, and the regeneration of the forest.

In the simplest case the long-term behavior of a dynamical system can be represented mathematically by the matrix equation

$$\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k, \quad \text{for } k = 0, 1, 2, \ldots,$$

where $\mathbf{A}$ is an $n \times n$ matrix, and $\mathbf{x}_k$ is an n element column vector whose elements describe the physical characteristics of the system at the kth stage. In a logging operation $n = 2$, and $\mathbf{x}_k = [T_k, R_k]^{\mathrm{T}}$, where T_k is the amount of timber remaining after k years and R_k is the amount of replanted timber that has matured after k years.

In general, let $\mathbf{A}$ be diagonalizable with the real eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$, and let the corresponding linearly independent eigenvectors be $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$. Then, if $\mathbf{x}_0$ describes the initial state of the system, since the eigenvectors form a basis for the system we may set $\mathbf{x}_0 = c_1\mathbf{u}_1 + c_2\mathbf{u}_2 + \ldots + c_n\mathbf{u}_n$. Use the representation of $\boldsymbol{x}_0$ to find a general expression for $\boldsymbol{x}_k$ in terms of the eigenvalues, and comment on the approximate form taken by $\boldsymbol{x}_k$ as k becomes large.

Given that

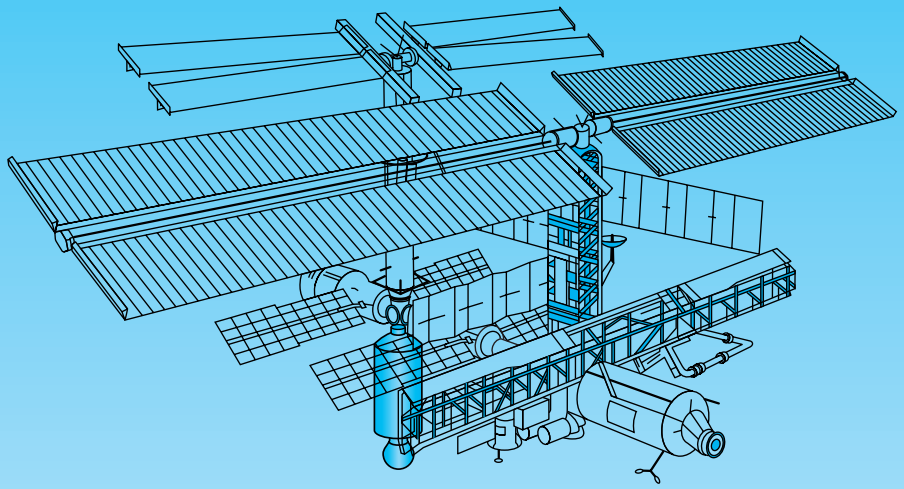
$$\mathbf{A} = \begin{bmatrix} 0.4 & 0.8 \\ -0.1 & 1.12 \end{bmatrix} \quad \text{and} \quad \mathbf{x}_k = \begin{bmatrix} T_k \\ R_k \end{bmatrix},$$

interpret the meaning of the coefficients of $\mathbf{A}$ in the context of a logging operation. Starting with $\boldsymbol{x}_0 = [1, 0.9]^{\mathrm{T}}$, generate the first 15 vectors $\boldsymbol{x}_k$, compare the results with the approximation found earlier, and comment on the result in terms of a logging operation.

Suggest a physical dynamical system where $\mathbf{A}$ is a 3×3 matrix. Define a suitable numerical matrix $\mathbf{A}$ and initial vector $\boldsymbol{x}_0$, generate the first 15 vectors $\boldsymbol{x}_k$, and interpret the results in terms of the model.

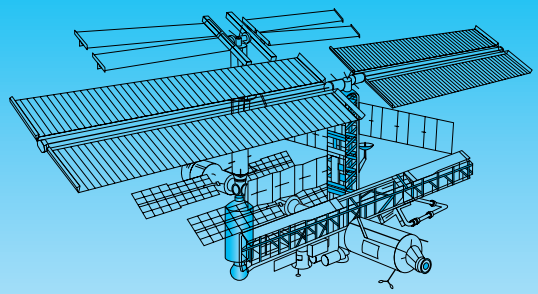
PART THREE

ORDINARY DIFFERENTIAL EQUATIONS



This Page Intentionally Left Blank

CHAPTER 5



First Order Differential Equations

Differential equations are fundamental to the study of engineering and physics, and this chapter marks the start of our discussion of this important topic. Typically, in an electrical problem, the dependent variable $i(t)$ in an ordinary differential equation might be the current flowing in a circuit at time t, in which case the independent variable would be the time. In all such examples, the nature of $i(t)$ depends on the current flow at the start, and the specification of information of this type is called an **initial condition** for the differential equation. Similarly, in chemical engineering, a dependent variable $m(t)$ might be the amount of a chemical produced by a reaction at time t. Here also the independent variable would be the time t, and to determine $m(t)$ in any particular case it would be necessary to specify the amount of $m(t)$ present at the start, that for convenience is usually taken to be when $t = 0$.

Many physical problems are capable of description in terms of a single first order ordinary differential equation, while other more complicated problems involve coupled first order differential equations, that after the elimination of all but one of the independent variables, can be replaced by a single higher order equation for the remaining dependent variable. This happens, for example, when determining the current in an R-L-C electrical circuit.

Thus first order ordinary differential equations can be considered as the building blocks in the study of higher order equations, and their properties are particularly important and easy to obtain when the equations are linear. The study and properties of the specially simple class of equations called **constant coefficient equations** is very important, as it forms the foundation of the study of higher order constant coefficient equations that will be developed later and have many and varied applications.

Motivation for the study of ordinary differential equations in general is provided by considering a number of typical problems that give rise to different types of differential equation. The first application involves the determination of orthogonal trajectories. A typical example of orthogonal trajectories arises in steady state two-dimensional temperature distributions, where one family of trajectories corresponds to the lines along which the temperature is constant, while the other family corresponds to lines along which heat flows. Other examples considered are the radioactive decay of a substance, the logistic equation and its connection with population growth, damped oscillations, the shape of a suspended power line, and the bending of beams.

The chapter starts by defining an ***m*th order** ordinary differential equation, of which a first order equation is a special case. Various important terms are defined, and the physical

significance of **initial** and **boundary conditions** for differential equations are introduced and explained.

The geometrical interpretation of the derivative dy/dx as the slope of a curve is used in Section 5.3 to develop the concept of the **direction field** associated with the first order equation $dy/dx = f(x, y)$. This concept is particularly useful as it leads to a geometrical picture showing the qualitative behavior of all solutions of the differential equation. It will be seen later that the idea underlying a direction field forms the basis of the simple Euler method for the numerical solution of an initial value problem.

First order equations are considered, **separable equations** are defined and solved, and some other special types of equation are introduced that arise in applications, of which the most important is the general **linear first order differential equation**. Its solution is found by using what is called an **integrating factor**. The first order linear differential equation is important, because the structure of its solution is typical of linear differential equations of all orders.

Another special first order equation that is considered is the Bernoulli equation. The Bernoulli equation is an important type of nonlinear equation with many applications, and in a sense it stands on the border between linear and nonlinear first order differential equations. An application of the Bernoulli equation is outlined in the text, and another more detailed one is to be found in the Exercise set at the end of Section 5.8.

The chapter ends by considering the important and practical questions concerning the existence and uniqueness of solutions of $dy/dx = f(x, y)$.

5.1 Background to Ordinary Differential Equations

An **ordinary differential equation (ODE)** is an equation that relates a function $y(x)$ to some of its derivatives $y^{(r)}(x) = d^r y/dx^r$. It is usual to call x the **independent** variable and y the **dependent variable**, and to write the most general ordinary differential equation as

$$F\left(x, y, y^{(1)}, y^{(2)}, \ldots, y^{(n)}\right) = 0. \tag{1}$$

The number n in (1) is called the **order** of the ordinary differential equation, and it is the order of the highest derivative of y that occurs in the equation. A class of ODEs of particular importance in engineering and science, because of their frequency of occurrence and the extensive analytical methods that are available for their solution, are the linear ordinary differential equations.

***n*th order linear variable coefficient equation**

The most general ***n*th order linear differential equation** can be written

$$a_0(x)\frac{d^n y}{dx^n} + a_1(x)\frac{d^{n-1} y}{dx^{n-1}} + \cdots + a_{n-1}(x)\frac{dy}{dx} + a_n(x)y = f(x), \tag{2}$$

with $a_0(x) \neq 0$ and we will consider it to be defined over some interval $a \leq x \leq b$. The functions $a_0(x), a_1(x), \ldots, a_n(x)$, called the **coefficients** of the equation, are known functions, and the known function $f(x)$ is called the **nonhomogeneous term**. The name **forcing function** is also sometimes given to $f(x)$, because in applications it represents the influence of an external input that drives a physical system represented by the differential equation. Equation (2) is called **homogeneous** if $f(x) \equiv 0$.

It will be seen later that the solution of the nonhomogeneous equation (2) is related in a fundamental manner to the solution of its associated homogeneous equation.

When one or more of the coefficients of (2) depend on x, it is called a **variable coefficient** equation. Simpler than **variable coefficient** linear equations, but still of considerable importance, are the linear equations in which the coefficients are the constants $a_0, a_1, \ldots, a_n$, so that (2) becomes

nth order linear constant coefficient equation

$$a_0\frac{d^n y}{dx^n} + a_1\frac{d^{n-1}y}{dx^{n-1}} + \cdots + a_{n-1}\frac{dy}{dx} + a_n y = f(x) \quad \text{for} \quad a \le x \le b. \tag{3}$$

Equations of this type are called **constant coefficient** linear equations.

If the interval $a \le x \le b$ on which equations (2) and (3) are defined is not specified, it is to be understood to be the largest one for which the equations have meaning. Sometimes, in the case of (2), this interval is determined by the variable coefficients $a_r(x)$, whereas in applications it is often determined by the nature of the problem that restricts x to a specific interval.

nonlinear equation and degree

An ordinary differential equation that is not linear is said to be **nonlinear**. Nonlinearity arises in ordinary differential equations because of the occurrence of a nonlinear function of the dependent variable y that sometimes occurs in the form of a power or a radical. The terms homogeneous and nonhomogeneous have no meaning for nonlinear equations.

A term that is also in use, mainly as an indication of the complexity to be expected of a solution, is the *degree* of an equation. The **degree** is the greatest power to which the highest order derivative in the differential equation is raised after the radicals have been cleared from expressions involving the dependent variable y.

EXAMPLE 5.1

(a) The ODE

$$\frac{dy}{dx} + 2xy = \sin x$$

is a linear variable coefficient nonhomogeneous first order equation.

(b) The ODE

$$(1 - x^2)\frac{d^2y}{dx^2} - 2x\frac{dy}{dx} + 6y = 0, \quad \text{with } -1 < x < 1,$$

is a linear variable coefficient homogeneous second order equation.

(c) The ODE

$$\frac{d^2y}{dx^2} + a\frac{dy}{dx} + by = \sin \omega x, \quad \text{with } \omega = \text{constant},$$

is a linear constant coefficient nonhomogeneous second order equation.

(d) The ODE

$$\frac{d^2\theta}{dt^2} + k \sin\theta = 0, \quad \text{with } k = \text{constant}$$

is a nonlinear second order equation because θ occurs nonlinearly in the function $\sin\theta$.

(e) The ODE

$$k\frac{d^2y}{dx^2} = f(x)[1 + (dy/dx)^2]^{3/2}, \quad \text{with } k > 0 \text{ a constant}$$

is a nonlinear second order equation of degree 2 involving a power and a radical. ■

general and particular solutions, and integral curves

A **solution** of an ordinary differential equation is a function $y = \Phi(x)$ that, when substituted into the equation, makes it identically zero over the interval on which the equation is defined. A solution of an nth order equation that contains n arbitrary constants is called the **general solution** of the equation. If the arbitrary constants in the general solution are assigned specific values, the result is called a **particular solution** of the equation.

singular solution

For obvious reasons the solution of an ordinary differential equation is also called an **integral curve**. A solution that cannot be obtained from the general solution for any choice of its arbitrary constants is called a **singular solution**. In the case of linear equations *all* possible solutions of the equation can be obtained from the general solution, so linear equations have no singular solutions. Nonlinear equations possess a more complicated structure that often allows the existence of one or more singular solutions.

EXAMPLE 5.2

(a) The general solution of the linear constant coefficient nonhomogeneous equation

$$\frac{d^2y}{dx^2} - 4y = x$$

is $y = Ae^{2x} + Be^{-2x} - x/4$, where A and B are arbitrary constants. This is easily checked, because substituting for y in the equation leads to the identity $x \equiv x$.

(b) The nonlinear equation

$$\left(\frac{dy}{dx}\right)^2 + y^2 = 1$$

has the general solution $y = \sin(x + A)$. However, $y = \pm 1$ are also seen to be solutions, though as these cannot be obtained from the general solution for any choice of A, they are *singular solutions*. ■

The linear equation (2) is often written in the more compact form

$$L[y] = f(x), \tag{4}$$

linear operator

where L is the **linear operator**

$$L[\cdot] \equiv a_0(x)\frac{d^n}{dx^n} + a_1(x)\frac{d^{n-1}}{dx^{n-1}} + \cdots + a_{n-1}(x)\frac{d}{dx} + a_n(x), \tag{5}$$

with coefficients that may or may not be functions of x. Only when $L[\cdot]$ acts on an n times differentiable function does it produce a function.

Equation (2) is called **linear** because if y_1 and y_2 are any two solutions of the homogeneous form of the equation $L[y] = 0$, the linear combination $y = C_1 y_1 + C_2 y_2$ where C_1 and C_2 are constants is also a solution. In terms of the differential operator $L[\cdot]$ this property becomes $L[C_1 y_1 + C_2 y_2] = C_1 L[y_1] + C_2 L[y_2]$, and it follows directly from the linearity of the differentiation operation, because

$$\frac{d^m}{dx^m}(y_1 + y_2) = \frac{d^m y_1}{dx^m} + \frac{d^m y_2}{dx^m},$$

for $m = 0, 1, \ldots, n$, with $d^0 y/dx^0 \equiv y$.

If $y_1(x), y_2(x), \ldots, y_m(x)$ are solutions of the nth order homogeneous equation $L[y] = 0$, with $m \leq n$ and $C_1, C_2, \ldots, C_m$ arbitrary constants, the linear combination

$$y(x) = C_1 y_1(x) + C_2 y(x) + \cdots + C_m y_m(x)$$

linear superposition

is called a **linear superposition** of the m solutions, and it is also a solution of the homogeneous equation.

Later we will define the linear independence of a set of functions over an interval and show that the homogeneous form of (2) has precisely n linearly independent solutions $y_1(x), y_2(x), \ldots, y_n(x)$, and that its general solution is

$$y_c(x) = C_1 y_1(x) + C_2 y(x) + \cdots + C_n y_n(x), \tag{6}$$

complementary solution, particular integral, and complete solution

where $C_1, C_2, \ldots, C_n$ are arbitrary constants. This general solution of the homogeneous form of equation (2) is called the **complementary function** or the **complementary solution** of (2). A function $y_p(x)$ that is a solution of the nonhomogeneous equation (2) but contains *no* arbitrary constants is called a **particular integral** of (2). The **complete solution** $y(x)$ of equation (2) is

$$y(x) = y_c(x) + y_p(x). \tag{7}$$

In applications of ordinary differential equations the values of the arbitrary constants in specific problems are obtained by choosing them so the solution satisfies auxiliary conditions that identify a particular problem.

Auxiliary conditions specified at a single point $x = a$, say, are called **initial conditions**, because x often represents the time so that conditions of this type describe how the solution starts. An **initial value problem (i.v.p.)** involves finding a solution of a differential equation that satisfies prescribed initial conditions.

boundary and initial conditions

A different type of problem arises when the auxiliary conditions are specified at two different points $x = a$ and $x = b$, say. Conditions of this type are called **boundary conditions**, because in such problems x usually represents a space variable, and the solution is required to be determined between two boundaries located at $x = a$ and $x = b$ where boundary conditions are prescribed. A **boundary value problem (b.v.p.)** involves finding a solution of a differential equation that satisfies prescribed boundary conditions.

EXAMPLE 5.3

(a) The linear nonhomogeneous ordinary differential equation

$$\frac{d^2y}{dx^2} + y = x$$

has the general solution $y = A\cos x + B\sin x + x$. This equation together with the initial conditions $y(0) = 0$, $y'(0) = 0$ specified at the point $x = 0$ constitutes an *initial value problem* for y. Choosing A and B to satisfy these initial conditions shows the unique solution of this i.v.p. to be $y = x - \sin x$ for $x \geq 0$.

(b) The linear homogeneous ordinary differential equation

$$\frac{d^2y}{dx^2} + y = 0$$

has the general solution $y = A\cos x + B\sin x$. This equation together with the conditions $y(0) = 0$, $y'(\pi/3) = 3$ specified at the two different points $x = 0$ and $x = \pi/3$ constitutes a *boundary value problem* for y. Choosing A and B to satisfy these conditions shows that this b.v.p. has the unique solution $y = 6\sin x$ for $0 < x < \pi/3$.

(c) Consider the linear homogeneous ordinary differential equation

$$\frac{d^2y}{dx^2} - y = 0 \quad \text{defined for } x \geq 0,$$

which is easily seen to have the general solution $y = Ae^x + Be^{-x}$. Imposing the boundary conditions $y(0) = 1$ and $y(+\infty) = 0$ constitutes a boundary value problem for y in which one condition is at $x = 0$ and the other is at plus infinity. The condition at infinity can only be satisfied if $A = 0$, so matching the solution $y = Be^{-x}$ to the condition $y(0) = 1$ shows that this b.v.p. has the **unique** (only) solution $y = e^{-x}$.

unique and nonunique solutions

(d) It is possible for a boundary value problem to have a unique solution as in (b), more than one solution, or no solution at all. More will be said about this later, but for the moment we give a simple example that shows why a boundary value problem may have many solutions or no solution.

The general solution of (b) is $y = A\cos x + B\sin x$, so if the boundary conditions $y(0) = 0$ and $y(\pi) = 0$ are imposed we find that $A = 0$ and B is indeterminate, so it may be assigned any value. In this case a solution certainly exists, as it is given by $y = B\sin x$, but B is arbitrary, so there is more than one solution. When more than one solution can be found that satisfies the auxiliary conditions, the solution is said to be **nonunique**.

If, in this example, the boundary conditions are replaced by $y(0) = 0$ and $y(\pi) = 1$, no choice of constants A and B can make the general solution satisfy the boundary conditions, so in this case there is no solution. ■

Summary

This section introduced the concept of an nth order ordinary differential equation, and the initial and boundary conditions that such equations are often required to satisfy. Emphasis was placed on linear equations and, in particular, on the structure of the solution of a linear first order equation, because the structure of the solution of this fundamental type of equation is shared by the solutions of all higher order linear equations.

EXERCISES 5.1

In Exercises 1 through 10, determine the order and degree of the equation and classify it as homogeneous linear, nonhomogeneous linear, or nonlinear.

1. $y''' + 3y'' + 4y' - y = 0$.
2. $y'' + 4y' + y = x \sin x$.
3. $y'' + x(y')^2 = \cosh x$.
4. $(y'')^{3/2} + xy' = [(1 + x)y']$.
5. $y'' + 3y' + 2y = x^2 \sin y$.
6. $y^{(4)} + x^2\sqrt{y} = 3 + x^3$.
7. $y' + 3xy = 1 + x^2$.
8. $y'' + y = \tan(y')$.
9. $(2 + x^2)y' + x(1 - y^2) = 0$.
10. $y'/y + \sin x = 3$.

5.2 Some Problems Leading to Ordinary Differential Equations

Before we develop methods for the solution of ordinary differential equations, it will be helpful to examine some simple geometrical and physical problems that lead to ODEs. There are many such problems, so we only consider some representative examples.

(a) A Geometrical Problem: Orthogonal Trajectories

The equation

$$F(x, y, c) = 0,$$

where the real variable c is a *parameter*, defines a **one-parameter** family of curves in the (x, y)-plane. This means that assigning a specific value to c determines a particular curve in the (x, y)-plane, and a different value of c will determine a different curve. It often happens that the equation $F(x, y, c) = 0$ defines y implicitly in terms of x, so that the equation cannot be solved explicitly as $y = f(x, c)$.

orthogonal trajectory

A curve that intersects every member of a one-parameter family of curves orthogonally (at right angles) is called an **orthogonal trajectory** of the family. A geometrical problem that often occurs is how to find a *family* of curves that form orthogonal trajectories to a given family.

isotherms, heat flow, streamlines, equipotentials, and flux lines

When some applications of conformal mapping to two-dimensional physical problems are considered in Chapter 17, it will be seen that orthogonal trajectories arise in the study of steady state heat conduction, fluid dynamics, and electromagnetic theory. In heat conduction (see Chapter 18), one family of curves represents lines of constant temperature called **isotherms**, and their orthogonal trajectories then represent **heat flow** lines. In two-dimensional fluid dynamics, orthogonal trajectories express the relationship between the curves followed by fluid particles called **streamlines**, and the associated **equipotential lines** along which a function called the **fluid potential** is constant. In two-dimensional electromagnetic theory an analogous situation arises where one family of curves describes lines of constant electric potential, again called **equipotential lines**, and the family of orthogonal trajectories that describes what are then called **flux lines**.

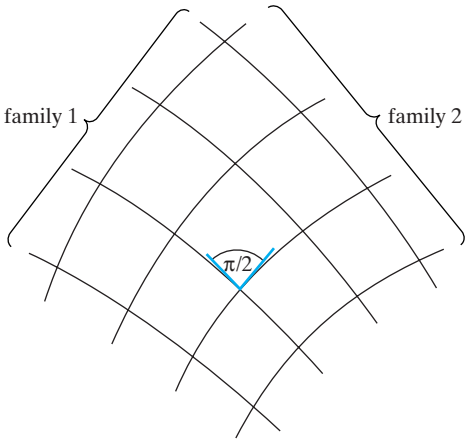


FIGURE 5.1 Two typical families of orthogonal trajectories.

Two typical families of orthogonal trajectories are illustrated in Fig. 5.1, and if these curves are related to steady state heat flow, family 1 could represent the isotherms and family 2 the heat flow lines.

Two specific examples of families of orthogonal trajectories are shown in Fig. 5.2, where in case (a) the curves are given by

$$x^2 + y^2 = c^2 \quad \text{and} \quad y = kx \qquad \text{(with } c \text{ and } k \text{ real)}.$$

The first equation describes a family of concentric circles centered on the origin, and the second family that forms their orthogonal trajectories comprises all the straight lines that pass through the origin.

In case (b) the curves are given by

$$x^2 - y^2 = c \quad \text{and} \quad xy = k \qquad \text{(with } c \text{ and } k \text{ real)},$$

where the two families of curves are families of mutually orthogonal rectangular hyperbolas.

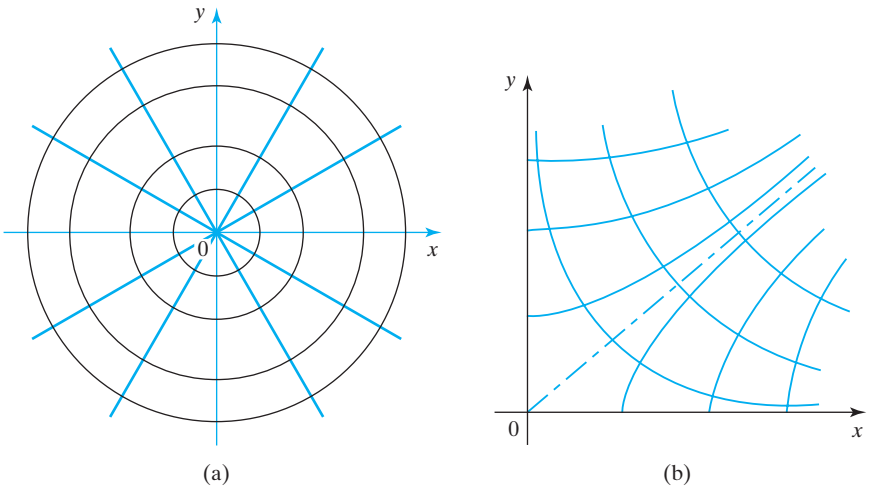


FIGURE 5.2 Specific examples of orthogonal trajectories.

In general the equation

$$F(x, y, c) = 0, \tag{8}$$

with c a parameter, describes a family of curves. To find their orthogonal trajectories we first need to obtain the differential equation for the family of curves determined by (8). This can be done by differentiating (8) with respect to x and then eliminating c between (8) and the equation with dy/dx to arrive at a differential equation of the form

$$\frac{dy}{dx} = f(x, y). \tag{9}$$

If the family of curves described by this differential equation is to be orthogonal to another family, the products of the gradients of every pair of intersecting curves must equal -1. So the gradient dy/dx of the family of curves that are mutually orthogonal to those of (9) must be such that

$$\frac{dy}{dx} = -\frac{1}{f(x, y)}. \tag{10}$$

This is the differential equation of the required family of orthogonal trajectories. In general (10) can often be solved by the method of separation of variables that will be discussed later.

(b) Chemical Reaction Rates and Radioactive Decay

In many circumstances, for a limited period of time, the rate of reaction of a chemical process can be considered to be proportional only to the amount Q of the chemical that is present at a given time t. The differential equation governing such a process then has the form

$$\frac{dQ}{dt} = kQ, \tag{11}$$

where $k \geq 0$ is a constant of proportionality. This is a homogeneous linear first order differential equation.

An analogous situation applies to the radioactive decay of an isotope for which the decay takes place at a rate proportional to the amount of radioactive isotope that is present at any given instant of time. The equation governing the amount Q of the isotope as a function of time t is also of the form shown in (11), but instead of the amount growing as in the previous case, it is decreasing, so as in this case the constant of proportionality is usually denoted by a positive number λ, the equation for radioactive decay takes the form

$$\frac{dQ}{dt} = -\lambda Q. \tag{12}$$

It is not difficult to see by inspection that the general solution of (12) is

$$Q = Q_0 e^{-\lambda t},$$

half-life

where Q_0 is the amount of the isotope present at the start when $t = 0$. The so-called **half-life** T_h of an isotope is the time taken for half of it to decay away, so setting $Q = (1/2)Q_0$ in the above result shows the half-life to be given by $T_h = (1/\lambda) \ln 2$.

(c) The Logistic Equation: Population Growth

In the study of phenomena involving the rate of increase of a quantity of interest, it often happens that the rate is influenced both by the amount of the quantity that is present at any given instant of time and by the limitation of a resource that is necessary to enable an increase to occur. Such a situation arises in a population of animals that compete for limited food resources, leading to the so-called *predator–prey* situations where an animal (the predator) feeds on another species (the prey) with the effect that overfeeding leads to starvation. This in turn leads to a reduction in the number of predators that in turn can lead to a recovery of the food stock. Similar situations arise in manufacturing when there is competition for scarce resources, and in a variety of similar situations.

To model the situation we let P represent the amount of the quantity of interest present at a given time t, and M represent the amount of resources available at the start. Then a simple model for this process is provided by the differential equation

$$\frac{dP}{dt} = kP(M - P), \tag{13}$$

logistic equation

in which k is a constant of proportionality. When constructing this equation the assumption has been made that the rate of increase dP/dt is proportional to both the amount P that is present at time t and to the amount $M - P$ that remains. Equation (13) is called the **logistic equation**, and it is nonlinear because of the presence of the term $-kP^2$ on the right, though it is easily integrated by the method of separation of variables to be described later.

(d) A Differential Equation that Models Damped Oscillations

Mechanical and electrical systems, and control systems in general, can exhibit oscillatory behavior that after an initial disturbance slowly decays to zero. The process producing the decay is a *dissipative* one that removes energy from the system, and it is called **damping**. To see the prototype equation that exhibits this phenomenon we need only consider the following very simple mechanical model. A mass M rests on a rough horizontal surface and is attached by a spring of negligible mass to a fixed point. The mass–spring system is caused to oscillate along the line of the spring by being displaced from its equilibrium position by a small amount and then released. Figure 5.3a shows the system in its equilibrium configuration, and Fig. 5.3b shows it when the mass has been displaced through a distance x from its rest position.

damping

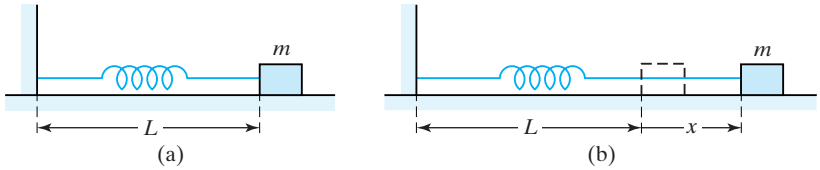


FIGURE 5.3 Mass–spring system.

If t is the time, the acceleration of the mass is d^2x/dt^2, so the force acting due to the motion is Md^2x/dt^2. The forces opposing the motion are the spring force, assumed to be proportional to the displacement x from the equilibrium position, and the frictional force, assumed to be proportional to the velocity dx/dt of the mass M. If the spring constant of proportionality is p and the frictional constant of proportionality is k, the two opposing forces are kdx/dt due to friction and px due to the spring. Equating the forces acting along the line of the spring and taking account of the fact that the spring and frictional forces oppose the force due to the acceleration shows the equation of motion to be the homogeneous second order linear equation

$$M\frac{d^2x}{dt^2} = -k\frac{dx}{dt} - px,$$

or

$$\frac{d^2x}{dt^2} + a\frac{dx}{dt} + bx = 0, \tag{14}$$

where $a = k/M$ and $b = p/M$.

If an external force $Mf(t)$ is applied to the spring, the equation governing the damped oscillations becomes the linear nonhomogeneous second order equation

$$\frac{d^2x}{dt^2} + a\frac{dx}{dt} + bx = f(t).$$

An equation of the same form as (14) governs the oscillation of the charge q in the R–L–C electric circuit shown in Fig. 5.4. The open circuit is shown in Fig. 5.4a with the plates of the capacitor C carrying initial charges Q and $-Q$, while Fig. 5.4b shows the circuit when the switch S has been closed, causing a current i to flow due to a charge is q at time t.

The respective potential drops in the direction of the arrow across the resistor R, the inductance L, and the capacitor C are $V = iR$, where $i = dq/dt$, Ldi/dt,

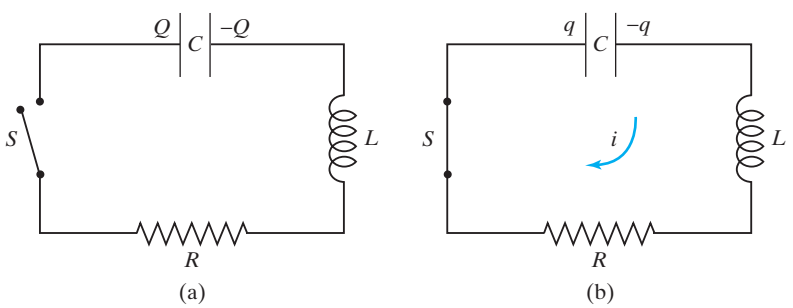


FIGURE 5.4 An R–L–C circuit.

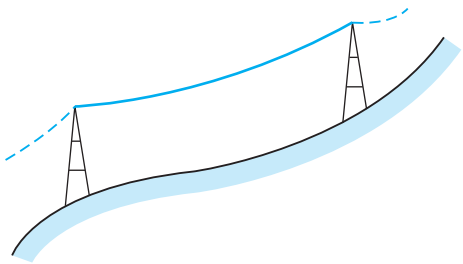


FIGURE 5.5 Suspended cable.

and q/C. Applying Kirchhoff's law, which requires the sum of the potential drops around the circuit to be zero, gives

$$L\frac{di}{dt} + Ri + \frac{q}{C} = 0.$$

Eliminating i by using the result $i = dq/dt$ leads to the following homogeneous linear second order equation for q:

$$LC\frac{d^2q}{dt^2} + RC\frac{dq}{dt} + q = 0.$$

This ODE is of the same form as (14) with $a = R/L$ and $b = 1/LC$.

(e) The Shape of a Suspended Power Line: The Catenary

An analysis of the forces acting on a power line attached to pylons as shown in Fig. 5.5, or on the suspension cable of a cable car, shows the shape of the cable to be determined by the solution $y(x)$ of the nonlinear differential equation

$$\frac{d^2y}{dx^2} = a\sqrt{1 + (dy/dx)^2}.$$

The shape taken by the cable is called a **catenary**, after the Latin word *catena*, meaning chain. Although this equation will not be solved here, it is not difficult to show that its solution is a hyperbolic cosine curve.

(f) Bending of Beams

An analysis of the forces and moments acting on a horizontal beam of uniform construction made from a material with Young's modulus E and supported at its two end points, with the moment of inertia of its cross-section about the central horizontal axis of the beam equal to I, leads to the following equation for the vertical

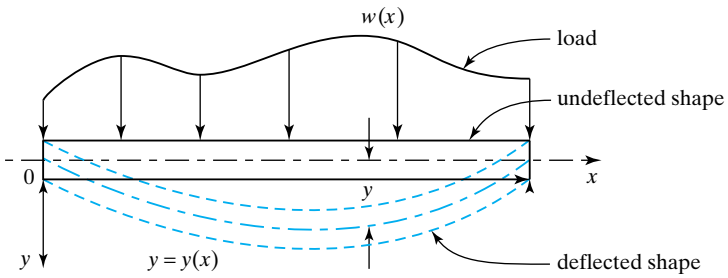


FIGURE 5.6 Deflection of a loaded beam.

deflection y caused by the weight of the beam and any loads it is supporting:

$$\frac{EI\, d^2y/dx^2}{[1+(dy/dx)^2]^{3/2}} = M(x). \tag{15}$$

Here $M(x)$ is the bending moment that acts to one side of a point x in the beam. If a distributed load of line density $w(x)$ acts along the beam creating a load $\int_a^b w(x)dx$ on the segment from $x = a$ to $x = b$, as represented in Fig. 5.6, it can be shown that $M(x)$ and $w(x)$ are related by the result

$$\frac{d^2M}{dx^2} = -w(x). \tag{16}$$

Using this result in (15) shows that the deflection $y(x)$ is determined by the solution of the nonlinear fourth order equation

$$\frac{d^2}{dx^2}\left\{\frac{EId^2y/dx^2}{[1+(dy/dx)^2]^{3/2}}\right\} = w(x), \tag{17}$$

flexural rigidity

in which the product EI is called the **flexural rigidity** of the beam. If the bending is small and the term $(dy/dx)^2$ can be neglected, (17) simplifies to the linear fourth order constant coefficient equation

$$\frac{d^4y}{dx^4} = \frac{w(x)}{EI},$$

which can be solved by direct integration.

Many applications of ordinary differential equations to physical problems are to be found in reference [3.6].

Summary

This section has provided mathematical and physical examples of problems that give rise to ordinary differential equations, some with initial conditions and others with boundary conditions. The logistic equation was seen to be nonlinear and first order, whereas others such as the equation governing radioactive decay and the equation describing damped

oscillations were seen to be linear and of first and second order, respectively. The beam equation is nonlinear, though when the bending is small it was seen to reduce to a simple linear fourth order equation that could be solved by direct integration.

EXERCISES 5.2

1. Derive the differential equation that describes the families of circles that are tangent to both the x- and y-axes.
2. Derive the differential equation satisfied by all curves such that the magnitude of the area under the curve between any two ordinates at $x = a$ and $x = b$ is proportional to the magnitude of the arc length of the curve from $x = a$ to $x = b$. Verify that the *catenary* $y(x) = k\cosh(x/k - K)$ is such a curve, with k and K parameters.

3.* A launch travels along the y-axis a constant speed U, starting from the origin, and a police launch starting from a point $a > 0$ on the x-axis pursues it at a constant speed $V > U$. If t is the time measured from the start of the pursuit, write down the differential equation that describes the pursuit path. At all times the police launch steers toward the first launch.

5.3 Direction Fields

In certain applications of mathematics it is necessary to know the qualitative behavior of solutions of a general first order equation

$$\frac{dy}{dx} = f(x, y) \tag{18}$$

over the entire (x, y)-plane, when either no analytical solution is available or, if one exists, it is too complicated to be useful. General properties of solutions of (18) that are known throughout the (x, y)-plane are called **global properties**. A typical global property might be that the solutions are known to be bounded for all x.

global properties

A numerical solution of (18) can always be obtained for any given initial condition (see Chapter 19), but it is impracticable to obtain such solutions for a large enough set of initial conditions simply to enable general the behavior of solutions all over the (x, y)-plane to be understood.

A convenient answer to this problem involves constructing a graphical representation of what is called the *direction field* of (18) at a conveniently chosen mesh of points covering a region R of interest in the (x, y)-plane.

The idea involved is simple and starts by dividing the interval $a \le x \le b$ into m subintervals of equal length $\Delta x = (b - a)/m$, and the interval $c \le y \le d$ into n subintervals of equal length $\Delta y = (d - c)/n$. The mesh of points to be used to cover R are then located at the points (x_r, y_s), where $x_r = a + r\Delta x$ and $y_s = c + s\Delta y$ with $r = 0, 1, \ldots, m$ and $s = 0, 1, \ldots, n$.

Once the mesh has been chosen, the function $f(x, y)$ is evaluated at each of the points (x_r, y_s). It follows directly that the number $f(x_r, y_s)$ associated with the point (x_r, y_s) is the *gradient* (slope) of the integral curve (solution curve) that passes through that point. Accordingly, the next step is to construct through each point (x_r, y_s), a small straight line segment making an angle $\theta_{rs} = \text{Arctan } f(x_r, y_s)$ with the x-axis, as in Fig. 5.7a.

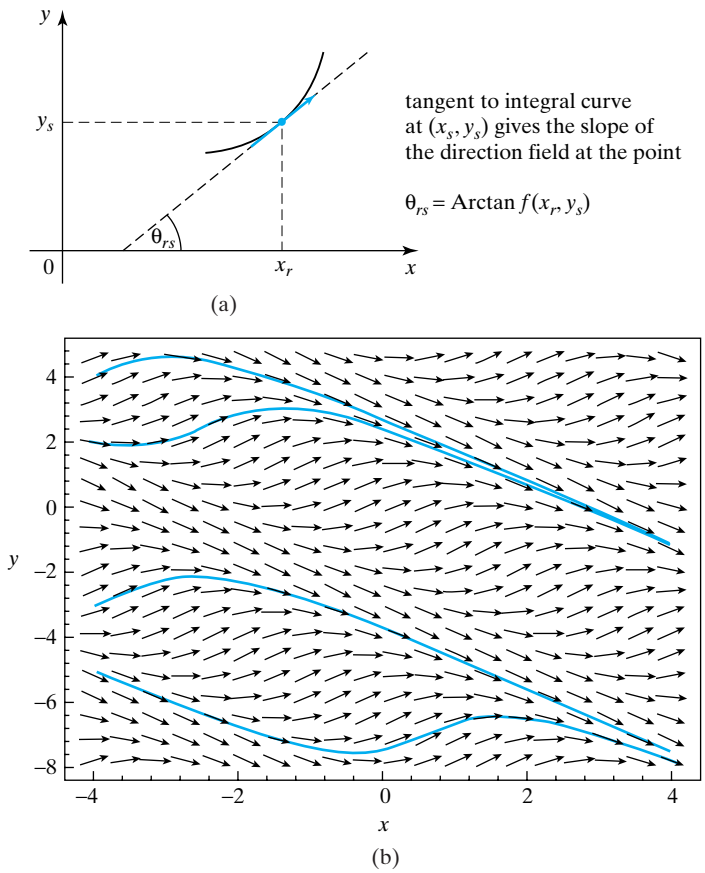


FIGURE 5.7 (a) The construction of a direction field vector at the point (x_r, y_s). (b) The direction field and integral curves for $dy/dx = \cos(x + y)$.

By the nature of their construction, each line segment that is drawn in this manner is tangent to the integral curve that passes through the point through which the segment is drawn. An examination of the pattern of the line segments indicates the overall pattern of behavior of all of the integral curves passing through region R. The assignment of a gradient $f(x, y)$ to each point of R is said to define the **direction field** of the ODE in (18) over R, and the method just described is its geometrical interpretation at a finite number of points of R.

direction field

The graphical interpretation of a direction field can be used to obtain an approximation to the integral curve that passes through an initial point (x_0, y_0) in R. This is accomplished by starting with the line segment through the point (x_0, y_0) and then joining up successive line segments as they intersect one another. As the construction of a direction field over a large region involves many calculations, it is usual to construct them with the aid of a computer.

The direction field for the nonlinear first order equation

$$\frac{dy}{dx} = \cos(x + y)$$

over the region $-4 \le x \le 4$ and $-8 \le y \le 5$ is shown in Fig. 5.7b, to which have been added some integral curves to show their relationship to the direction field.

Summary

The concept of a direction field of a first order differential equation $dy/dx = f(x, y)$ was introduced in this section. It is a graphical representation of the slope (gradient) of solution curves of the differential equation where they pass through a rectangular mesh of points inside a region of the (x, y)-plane where the solution of the differential equation is of interest. It involves plotting at each mesh point (x_i, y_i) a short segment of the tangent to the solution curve with slope $f(x_i, y_i)$ that passes through that point, to which is added an arrow showing the direction in which the solution is changing as x increases. A direction field provides a geometrical representation of the global nature of the solution inside the region of interest, and tracing successive line segments from one to another, starting from any mesh point, provides a rough picture of the solution curve that originates from the initial condition represented by that mesh point.

EXERCISES 5.3

In each of the following exercises, with the aid of a computer algebra package: (a) Construct the direction field for the given equation at a suitable number of mesh points, (b) use the results of (a) to sketch some representative integral curves, and (c) compare an approximate integral curve through a chosen initial point (x_0, y_0) with the exact solution found by requiring the given general solution to pass through that point.

1. $dy/dx = y + 2x$; $y = Ce^x - 2 - 2x$.
2. $dy/dx = y + 2\cos x$; $y = Ce^x - \cos x + \sin x$.
3. $dy/dx = 2x - y$; $y = Ce^{-x} - 2 + 2x$.
4. $dy/dx = x(1 + y/2)$; $y = C\exp(x^2/4) - 2$.
5. $dy/dx = y + x^2$; $y = Ce^x - 2 - 2x - x^2$.

5.4 Separable Equations

Sometimes the function $f(x, y)$ in the first order differential equation

$$\frac{dy}{dx} = f(x, y) \tag{19}$$

can be written as the product of a function $F(x)$ depending only on x and a function $G(y)$ depending only on y, so that $f(x, y) = F(x)G(y)$, allowing (19) to be written

$$\frac{dy}{dx} = F(x)G(y). \tag{20}$$

two forms of a separable equation

When (19) can be expressed in this simple form, its variables x and y are said to be **separable**, and the equation itself to be of **variables separable** type. If we use differential notation, (20) becomes

$$\frac{1}{G(y)}dy = F(x)dx, \tag{21}$$

so provided $G(y) \neq 0$, equation (21) can be solved by routine integration of the left side with respect to y and of the right side with respect to x. Thus, in principle, the solution of a first order differential equation in which the variables are separable can always be found, though in practice the integrals involved may be difficult or sometimes impossible to evaluate analytically.

Separable first order equations

The differential equation

$$\frac{dy}{dx} = f(x, y)$$

is said to be **separable** if it can be written in the form

$$\frac{dy}{dx} = F(x)G(y),$$

or, in differential form,

$$\frac{1}{G(y)}dy = F(x)dx.$$

EXAMPLE 5.4

examples of separable equations

Solve the logistic equation

$$\frac{dP}{dt} = kP(M - P)$$

given in equation (13) of Section 5.2(c), assuming $k > 0$ and $0 \le P \le M$. Find the solution of the initial value problem in which $P = P_0$ when $t = 0$, and draw some typical integral curves.

Solution The equation is separable and can be written in the differential form

$$\frac{dP}{P(M - P)} = kdt.$$

If we write the left-hand side in partial fraction form, the equation becomes

$$\frac{dP}{P} + \frac{dP}{(M - P)} = Mkdt,$$

and after integration we find that

$$\ln\left|\frac{P}{M - P}\right| = Mkt + C,$$

where C is an arbitrary constant of integration. As the solution for P must lie in the interval $0 \le P \le M$, this result simplifies to

$$P = \frac{MA}{A + \exp(-Mkt)},$$

where A is an arbitrary constant.

The arbitrary constant A is related to C by $A = e^C$, but as C is arbitrary, the constant A is also arbitrary, so for simplicity we denote the arbitrary constant in this last result by A without mentioning how it is related to C. In general, arithmetic is not usually performed on arbitrary constants, so after algebraic manipulations, either constants are renamed or the same symbol is used for a related constant.

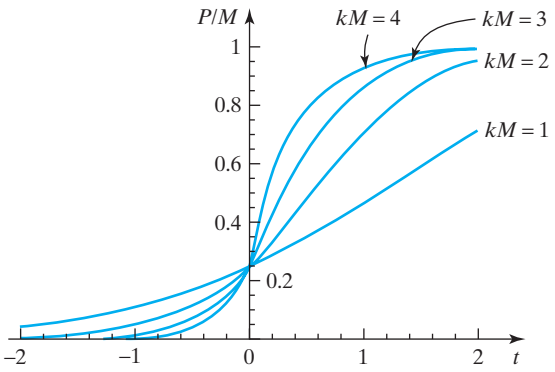


FIGURE 5.8 Integral curves for the logistic equation.

To solve the initial value problem we must find A such that $P = P_0$ when $t = 0$, from which it is easily seen that $A = P_0/(M - P_0)$. The required particular solution is thus

$$P = \frac{MP_0}{P_0 + (M - P_0)\exp(-Mkt)}.$$

Representative integral curves of $P(t)/M$ obtained from this expression using $P_0/M = \frac{1}{4}$ and $kM = 1, 2, 3$, and 4 are shown in Fig. 5.8 for $-2 \le t \le 2$. ■

EXAMPLE 5.5 Solve the initial value problem for the equation expressed in differential form

$$x^2y^2dx - (1 + x^2)dy = 0, \quad \text{given that } y(0) = 1.$$

Solution The equation is separable because it can be written

$$\frac{dy}{y^2} = \frac{x^2}{(1 + x^2)}dx.$$

Integration gives

$$\int \frac{dy}{y^2} = \int \frac{x^2}{(1 + x^2)}dx,$$

and after the integrations have been performed this becomes

$$-1/y = x - \text{Arctan}\, x + C,$$

where C is an arbitrary constant of integration. This general solution will satisfy the initial condition $y(0) = 1$ if $C = -1$, so the required solution is seen to be

$$y = 1/(\text{Arctan}\, x - x + 1).$$ ■

EXAMPLE 5.6 Derive the differential equation that determines the orthogonal trajectories of the one parameter family of curves $y = Cxe^x$, and solve it to find the equation of these trajectories.

Solution The differential equation describing the family of curves $y = Cxe^x$ is found by first calculating $y'(x)$, and then using the original equation to eliminate C

from the result. We have

$$y'(x) = Ce^x(1+x),$$

but from the original equation $C = y/xe^x$, so eliminating C between these two results shows that the required differential is

$$y'(x) = y(1+x)/x.$$

The product of the gradient $y'(x)$ of curves belonging to this family and the gradient of the family of orthogonal trajectories must equal -1 (see Section 5.2(a)), so the differential equation of the orthogonal trajectories is the separable equation

$$\frac{dy}{dx} = -\frac{x}{y(1+x)}.$$

After separation of the variables and integration, this becomes

$$\int y dy = -\int \frac{x}{1+x} dx,$$

so that

$$y^2 = \ln(1+x)^2 - 2x + C.$$ ■

EXAMPLE 5.7

A circular metal radiator pipe has inner radius R_1 and outer radius R_2 ($R_2 > R_1$). When operating under steady conditions the radial temperature distribution $T(r)$ in the metal wall of the pipe is known to be a solution of the ordinary differential equation (see the heat equation in cylindrical polar coordinates in Section 18.5)

$$r\frac{d^2T}{dr^2} + \frac{dT}{dr} = 0.$$

(i) Find the radial temperature distribution in the pipe wall when the inner surface is maintained at a constant temperature T_1 and the outer surface is maintained at a constant temperature T_2.

(ii) Find the radial temperature distribution in the pipe wall when the inner surface is maintained at a constant temperature T_1 and heat is lost by radiation from the outer surface according to Newton's law of cooling that requires the heat flux across the outer surface to be proportional to the difference in temperature between the surface and the surrounding air at a temperature T_2.

Solution

(i) Setting $u = dT/dr$ the equation becomes the separable equation

$$r\frac{du}{dr} + u = 0 \quad \text{and so} \quad \frac{du}{u} = -\frac{dr}{r},$$

from which it follows that

$$\ln u = -\ln r + \ln A,$$

where for convenience the arbitrary integration constant has been written $\ln A$. Thus $ur = A$, so after substituting for u and again separating variables we have

$$\frac{dT}{dr} = \frac{A}{r}.$$

A final integration gives the general solution

$$T(r) = A \ln r + B,$$

where B is another arbitrary integration constant.

Matching the arbitrary constants A and B to the required conditions $T(R_1) = T_1$ and $T(R_2) = T_2$ then gives the required solution

$$T(r) = \frac{T_1 \ln(R_2/r) + T_2 \ln(r/R_1)}{\ln(R_2/R_1)}.$$

(ii) The heat flux across the surface $r = R_2$ is proportional to dT/dr at $r = R_2$, and this in turn is proportional to the temperature difference $T(R_2) - T_2$, so the required boundary condition on the outer surface of the pipe is of the form

$$\left(\frac{dT}{dr}\right)_{r=R_2} = -h[T(R_2) - T_2],$$

where the negative sign is necessary because heat is being lost across the surface $r = R_2$, and h is a constant depending on the metal in the pipe and the heat transfer condition at its surface.

The general solution is still $T(r) = A \ln r + b$, but now the arbitrary constants A and B must be matched to the condition $T(R_1) = T_1$ on the inside wall of the pipe, and to the above condition derived from Newton's law of cooling. When this is done the temperature distribution in the pipe is found to be

$$T(r) = T_1 + \frac{hR_2(T_2 - T_1)}{1 + hR_2 \ln(R_2/R_1)} \ln\left(\frac{r}{R_1}\right). \quad \blacksquare$$

Summary

This section introduced the important class of separable differential equations $dy/dx = F(x)G(y)$, so called because when written in the form $dy/G(y) = F(x)dx$ the variables are separated by the = sign; they can be integrated immediately provided antiderivatives (indefinite integrals) of $1/G(y)$ and $F(x)$ can be found. This method was used to integrate the nonlinear logistic equation and to obtain the equation of some orthogonal trajectories.

EXERCISES 5.4

In Exercises 1 through 4 solve the given differential equation by hand and confirm the result by using computer algebra.

1. $2yy' = x(1 - 2y)$ with $y(1) = 1$.
2. $2x^2y^2y' + y^4 = 4$ with $y(1) = 3$.
3. $(x^2 - 4)y' = x(1 - 2y)$ with $y(\sqrt{5}) = 1$.
4. $2\sqrt{(1 + x^2)}y' = \sqrt{(1 - y^2)}$ with $y(1) = 1$.

In Exercises 5 through 14 find the general solution of the given differential equation.

5. $\sqrt{(1 + x^2)}y' - 3x\sqrt{(y^2 - 1)} = 0$.
6. $e^{-3x}y' + x \sin 2y = 0$.
7. $2(1 + x)(1 + y)y' + (y + 2)^2 = 0$.
8. $2(x - 1)y' + (x^2 - 2x + 3) \cos^2 y = 0$.
9. $(1 + 3y^2)y' + 2y \ln|1 + x| = 0$.
10. $2(1 - \cos x)y' + 3 \sin y = 0$.
11. $(1 + x^2)yy' - x(y^2 + y + 1) = 0$.
12. $(x^2 + 9)y^2y' - \sqrt{(4 - y^2)} = 0$.
13. $y' \operatorname{ctg} x + 2y = 4$.
14. $(x + 1)y^2y' = x(y^2 + 4)$.

In Exercises 15 through 17 derive and then solve the differential equation that determines the orthogonal trajectories to the given one parameter family of curves.

15. $y = b + k(x - a)$ with a and b constants and k a parameter.

16. $x^2 - 4y^2 + y = c$ with c a parameter.

17. $y = Cx^2e^{2x}$ with C a parameter.

18. A snowball of radius 2 inches is brought into a warm room at a constant temperature above freezing point, and it is found that after 6 hours it has melted to a radius of 1.5 inches. Assuming the melting occurs at a rate proportional to the surface area, write down the differential equation determining the radius as a function of time t in hours, and find the general expression for the radius as a function of time. Comment on any deficiency exhibited by this mathematical model.

19. A simple model called *Malthus' law* for the change in a bacterial population $N(t)$ as a function of time t involves assuming the rate of change is proportional to the population present at time t. Write down the differential equation governing $N(t)$ if the constant of proportionality is $\lambda > 0$, and find an expression for $N(t)$ given that initially $N(0) = N_0$. Find λ if $N(t_1) = N_1$ when $t = t_1$ and $N(t_2) = N_2$ when $t = t_2$, with $N_1 > N_2$ and $t_2 > t_1$. Give a reason why this model is unrealistic when t is large.

20. When a beam of light enters a parallel slab of transparent material at right angles to its plane surface, its intensity I decreases at a rate proportional to the intensity $I(x)$ at a perpendicular distance x into the material. Given a slab of material where the intensity at a distance h into the slab is 40% of the initial intensity, write down the differential equation for $I(x)$. Solve the equation for $I(x)$ and find the distance at which the intensity is 10% of its initial value.

21. The dating of a fossilized bone is based on the amount of radioactive isotope carbon-14 present in the bone. The method uses the fact that the isotope is produced in the atmosphere at a steady rate by bombardment of nitrogen by cosmic radiation when it is absorbed into the living bone. The process stops when the bone is dead, after which the C-14 present in the bone decays exponentially. Assuming the half-life of C-14 is 5600 years, and a bone is found to contain 1/500*th* of the original amount of C-14 that was present originally, determine its age. This approach is called **radioactive carbon dating**.

22. A cylindrical tank of cross-sectional area A standing in a vertical position is filled with water to a depth h. At time $t = 0$ a circular hole of radius a in the bottom of the tank is opened and water is allowed to drain away under gravity. It is known from *Torricelli's law* that the speed of flow of the water through the hole when the water in the tank has depth x is equal to $\sqrt{2gx}$, this being the speed attained by a particle falling freely from rest under gravity through a distance x, where g is the acceleration due to gravity. Write down the differential equation determining the water height $x(t)$ in the tank when $t > 0$, and solve the equation for $x(t)$. If water is added to the tank at a rate $V(t)$, write down the modified equation governing the water height. If $V(t) = V_0$ is constant, and the flow into and out of the tank reaches equilibrium, find the equilibrium height of the water in the tank. Remark: In applications the expression $\sqrt{2gx}$ is replaced by $k\sqrt{2gx}$, with $0 < k < 1$ a constant. The factor k allows for the contraction of the jet after leaving the hole. In the case of water $k \approx 0.6$.

5.5 Homogeneous Equations

homogeneous equation of degree *n*

A function $f(x, y)$ is said to be **algebraically homogeneous of degree *n***, or simply **homogeneous of degree *n***, if $f(tx, ty) = t^n f(x, y)$ for some real number n and all $t > 0$, for $(x, y) \neq (0, 0)$.

EXAMPLE 5.8

(a) If $f(x, y) = x^2 + 3xy + 4y^2$, then $f(tx, ty) = t^2(x^2 + 3xy + 4y^2) = t^2 f(x, y)$, so $f(x, y)$ is homogeneous of degree 2.

(b) If $f(x, y) = \ln|y| - \ln|x|$ for $(x, y) \neq (0, 0)$, then $f(x, y) = \ln|y/x|$, so $f(tx, ty) = f(x, y)$, showing that $f(x, y)$ is homogeneous of degree 0.

(c) If

$$f(x, y) = \frac{x^{3/2} + x^{1/2}y + 3y^{3/2}}{2x^{3/2} - xy^{1/2}}, \text{ then } f(tx, ty) = t^0 f(x, y),$$

showing that $f(x, y)$ is homogeneous of degree 0.

(d) If

$$f(x, y) = x^2 + 4y^2 + \sin(x/y), \text{ then } f(tx, ty) = t^2(x^2 + 4y^2) + \sin(x/y),$$

so $f(x, y)$ is *not* homogeneous, because although both the first group of terms and the last term are homogeneous functions of x and y, they are not both homogeneous of the same degree.

(e) If $f(x, y) = \tan(xy + 1)$, then $f(tx, ty) = \tan(t^2xy + 1)$, so $f(x, y)$ is *not* homogeneous. ■

Homogeneous differential equations

The first order ODE in differential form

$$P(x, y)dx + Q(x, y)dy = 0$$

is called **homogeneous** if P and Q are homogeneous functions of the same degree or, equivalently, if when written in the form

$$\frac{dy}{dx} = f(x, y),$$

the function $f(x, y)$ can be written as

$$f(x, y) = g(y/x).$$

The substitution $y = ux$ will reduce either form of the homogeneous equation to an equation involving the independent variable x and the new dependent variable u in which the variables are separable. As with most separable equations the solution can be complicated, and it is often the case that y is determined implicitly in terms of x.

EXAMPLE 5.9 Solve

$$(y^2 + 2xy)dx - x^2\,dy = 0.$$

Solution Both terms in the differential equation are homogeneous of degree 2, so the equation itself is homogeneous. Differentiating the substitution $y = ux$ gives

$$\frac{dy}{dx} = u + x\frac{du}{dx}, \quad \text{or} \quad dy = udx + xdu.$$

After substituting for y and dy in the differential equation and cancelling x^2, we obtain the variables separable equation

$$u(u + 1)dx = xdu, \quad \text{or} \quad \frac{du}{u(u + 1)} = \frac{dx}{x}.$$

This has the general solution

$$u = \frac{Cx}{1 - Cx}, \quad \text{but} \quad y = ux \quad \text{and so } y = \frac{Cx^2}{1 - Cx},$$

where C is an arbitrary constant. In this case the general solution is simple and y is determined explicitly in terms of x. ■

EXAMPLE 5.10 Solve

$$\frac{dy}{dx} = \frac{y^2}{xy - x^2}.$$

Solution The equation is homogeneous because it can be written

$$\frac{dy}{dx} = \frac{(y/x)^2}{(y/x) - 1}.$$

Making the substitution $y = ux$, and again using the result $dy/dx = u + xdu/dx$, reduces this to the separable equation

$$u + x\frac{du}{dx} = \frac{u^2}{u - 1}, \text{ or } \left(1 - \frac{1}{u}\right)du = \frac{dx}{x}.$$

Integration gives

$$u - \ln|u| = \ln|x| + \ln|C|,$$

where C is an arbitrary integration constant. Finally, substituting $u = y/x$ and simplifying the result we arrive at the following implicit solution for y:

$$y = Ce^{y/x}.$$

■

An equation of the form

$$\frac{dy}{dx} = \frac{ax + by + c}{px + qy + r}$$

near-homogeneous

is called **near-homogeneous**, because it can be transformed into a homogeneous equation by means of a variable change that shifts the origin to the point of intersection of the two lines

$$ax + by + c = 0 \quad \text{and} \quad px + qy + r = 0.$$

EXAMPLE 5.11 Solve the initial value problem

$$\frac{dy}{dx} = \frac{y + 1}{x + 2y} \quad \text{with } y(2) = 0.$$

Solution The equation is near-homogeneous and the lines $y + 1 = 0$ and $x + 2y = 0$ intersect at the point $x = 2$ and $y = -1$, so we make the variable change $x = X + 2$ and $y = Y - 1$, as a result of which the equation becomes the homogeneous equation

$$\frac{dY}{dX} = \frac{Y}{X + 2Y}.$$

Solving this as in Example 5.9 by setting $Y = uX$ leads to the equation

$$-\left(\frac{1 + 2u}{2u^2}\right)du = \frac{dX}{X},$$

with the solution

$$1/u = 2\ln|CuX|,$$

where C is an arbitrary integration constant. If we set $u = Y/X$, this becomes

$$X = 2Y \ln |CY|,$$

where C is an arbitrary constant. Returning to the original variables by substituting $X = x - 2$, $Y = y + 1$, we arrive at the required general solution

$$x = 2 + 2(y + 1) \ln |C(y + 1)|.$$

Although this is an implicit solution for y, if we regard y as the independent variable and x as the dependent variable, solution curves (integral curves) are easily graphed. Substituting the initial condition $y = 0$ when $x = 2$ in the general solution shows that $C = 1$, so the solution of the initial value problem is

$$x = 2 + 2(y + 1) \ln |y + 1|. \quad \blacksquare$$

Summary

This section introduced the special type of first order ordinary differential equation known as an algebraically homogeneous equation. This name is frequently shortened to the term homogeneous equation, though this must not be confused with the sense in which the term homogeneous is used in Section 5.1. After showing how such equations can be solved, it was shown how a simple linear change of variables changes a near-homogeneous equation to a homogeneous equation that can then be solved.

EXERCISES 5.5

In Exercises 1 through 14 find by hand calculation the general solution of the given homogeneous or near-homogeneous equations and confirm the result by using computer algebra.

1. $y' = y/(2x + y)$.
2. $y' = (2xy + y^2)/(3x^2)$.
3. $y' = (2x^2 + y^2)/xy$.
4. $y' = (2xy + y^2)/x^2$.
5. $y' = (x - y)/(x + 2y)$.
6. $y' = (x + 4y)/x$.
7. $y' = (2x + y \cos^2(y/x))/(x \cos^2(y/x))$.
8. $y' = 3y^2/(1 + x^2)$.
9. $y' = (x + y \sin^2(y/x))/(x \sin^2(y/x))$.
10. $y' = 3x \exp(x + 2y)/y$.
11. $y' = (y + 2)/(x + y + 2)$.
12. $y' = (y + 1)/(x + 2y + 2)$.
13. $y' = (x + y + 1)/(x - y + 1)$.
14. $y' = (x - y + 1)/(x + y)$.

5.6 Exact Equations

The so-called *exact* equations have a simple structure, and they arise in many important applications as, for example, in the study of thermodynamics. After definition of an exact equation, a test for exactness will be derived and the general solution of such an equation will be found.

Exact equations

definition of an exact equation

The first order ODE

$$M(x, y)dx + N(x, y)dy = 0$$

is said to be **exact** if a function $F(x, y)$ exists such that the total differential

$$d[F(x, y)] = M(x, y)dx + N(x, y)dy.$$

It follows directly that if

$$M(x, y)dx + N(x, y)dy = 0 \tag{22}$$

is exact, then the total differential

$$d[F(x, y)] = 0,$$

so the general solution of (22) must be

$$F(x, y) = \text{constant}. \tag{23}$$

EXAMPLE 5.12

The total differential of $F(x, y) = 3x^3 + 2xy^2 + 4y^3 + 2x$ is

$$\begin{aligned} d[F(x, y)] &= (\partial F/\partial x)dx + (\partial F/\partial y)dy \\ &= (9x^2 + 2y^2 + 2)dx + (4xy + 12y^2)dy, \end{aligned}$$

so the exact differential equation

$$(9x^2 + 2y^2 + 2)dx + (4xy + 12y^2)dy = 0$$

has the general solution

$$3x^3 + 2xy^2 + 4y^3 + 2x = \text{constant}.$$

■

Three questions now arise:

(i) Is there a test for exactness?
(ii) If an equation is exact, is it possible to find its general solution?
(iii) If an equation is not exact, is it possible to modify it to make it exact?

There are satisfactory answers to the first two questions, and a less satisfactory answer to the third question. We deal with the last question first.

It can be shown that an equation of the form (21) that is *not* exact can always be made exact if it is multiplied by a suitable factor $\mu(x, y)$, called an **integrating factor**, though there is no general method by which such an integrating factor can be found. Fortunately, however, an integrating factor can always be found for a variable coefficient linear first order ODE, and in the next section the integrating factor will be derived for such an ODE and then used to find its general solution.

We now turn our attention to the first question. If $F(x, y) = \text{constant}$ is a solution of the exact differential equation

$$M(x, y)dx + N(x, y)dy = 0, \tag{24}$$

then $M(x, y) = \partial F/\partial x$ and $N(x, y) = \partial F/\partial y$. So, provided the derivatives $\partial F/\partial x$, $\partial F/\partial y$, $\partial^2 F/\partial x \partial y$, and $\partial^2 F/\partial y \partial x$ are defined and continuous in the region within which the differential equation is defined, the mixed derivatives will be equal so that $\partial^2 F/\partial x \partial y = \partial^2 F/\partial y \partial x$. This last result is equivalent to requiring that $\partial M/\partial y = \partial N/\partial x$ in order that (24) is exact, so this provides the required test for exactness.

THEOREM 5.1

Test for exactness The differential equation

$$M(x, y)dx + N(x, y)dy = 0$$

a simple test for exactness

is exact if and only if $\partial M/\partial y = \partial N/\partial x$. ■

EXAMPLE 5.13

Test for exactness the differential equations

(a) $\{\sin(xy + 1) + xy\cos(xy + 1)\}dx + x^2\cos(xy + 1)dy = 0.$

(b) $(2x + \sin y)dx + (2x\cos y + y)dy = 0.$

Solution In case (a) $M(x, y) = \sin(xy + 1) + xy\cos(xy + 1)$ and $N(x, y) = x^2\cos(xy + 1)$, and $\partial M/\partial y = \partial N/\partial x$, so the equation is exact.

In case (b) $M(x, y) = 2x + \sin y$ and $N(x, y) = 2x\cos y + y$ but $\partial M/\partial y \neq \partial N/\partial x$, so the equation is not exact. ■

Having established a test for exactness, it remains for us to determine how the general solution of an exact equation can be found. The starting point is the fact that if $F(x, y) =$ constant is a solution of the exact equation

$$M(x, y)dx + N(x, y)dy = 0,$$

then $\partial F/\partial x = M(x, y)$ and $\partial F/\partial y = N(x, y)$.

Two expressions for $F(x, y)$ can be obtained from these results by integrating M with respect to x while regarding y as a constant, and integrating N with respect to y while regarding x as a constant, because this reverses the process of partial differentiation by which M and N were obtained. However, after integrating M it will be necessary to add not only an arbitrary constant, but also an arbitrary function $f(y)$ of y, because this will behave like a constant when F is differentiated partially with respect to x to obtain M. Similarly, after integrating N it will be necessary to add not only an arbitrary constant, but also an arbitrary function $g(x)$ of x, because this will behave like a constant when F is differentiated partially with respect to y to obtain N.

These two expressions for F will look different but must, of course, be identical. The arbitrary function $f(y)$ can be found by identifying it with any function only of y that occurs in the expression for F obtained by integrating N, while the arbitrary function $g(x)$ can be found by identifying it with any function only of x that occurs in the expression for F found by integrating M, where, of course, the true constants introduced after each integration must be identical.

EXAMPLE 5.14

Show the following equation is exact and find its general solution:

$$\{3x^2 + 2y + 2\cosh(2x + 3y)\}dx + \{2x + 2y + 3\cosh(2x + 3y)\}dy = 0.$$

Solution In this equation $M(x, y) = 3x^2 + 2y + 2\cosh(2x + 3y)$, and $N(x, y) = 2x + 2y + 3\cosh(2x + 3y)$, so as $M_y = N_x = 2 + 6\sinh(2x + 3y)$ the equation is exact:

$$F(x, y) = \int M(x, y)dx = \int \{3x^2 + 2y + 2\cosh(2x + 3y)\}dx$$
$$= x^3 + 2xy + \sinh(2x + 3y) + f(y) + C,$$

and

$$F(x, y) = \int N(x, y)dy = \int \{2x + 2y + 3\cosh(2x + 3y)\}dy$$
$$= 2xy + y^2 + \sinh(2x + 3y) + g(x) + D.$$

For these two expressions to be identical, we must set $f(y) \equiv y^2$, $g(x) \equiv x^3$, and $D = \mathrm{C}$, so $F(x, y)$ is seen to be

$$F(x, y) = x^3 + 2xy + y^2 + \sinh(2x + 3y) + C,$$

and so the general solution is

$$x^3 + 2xy + y^2 + \sinh(2x + 3y) = C,$$

where as C is an arbitrary constant we have chosen to write C rather than $-C$ on the right of the solution. ■

Summary

This section introduced the class of first order ordinary differential equations known as exact equations that arise in many different applications. It was then shown how the equality of mixed derivatives yields a simple test for exactness.

EXERCISES 5.6

In Exercises 1 through 8 test the equation for exactness, and when an equation is exact, find its general solution.

1. **(a)** $\{\sin(3y) + 4x^2y\}dx + \{3x\cos(3y) + y + 2x^3\}dy = 0;$
(b) $\{4x^3 + 3y^2 + \cos x\}dx + \{6xy + 2\}dy = 0.$

2. **(a)** $\{(2x + 3y^2)^{-1/2} + 4y^3 + 2x\}dx + \{3y/(2x + 3y^2) + 12xy^2\}dy = 0;$
(b) $\{\cos(x + 3y^2) + 4xy^3\}dx + \{6y\cos(x + 3y^2) + 3x^2y^2 + 2y\}dy = 0.$

3. **(a)** $\{\sin x + x\cos x + \cosh(x + 2y)\}dx + \{3y^2 + 2\cosh(x + 2y)\}dy = 0;$
(b) $\{6x(2x^2 + y^2)^{1/2} + x^2\}dx + 2y(2x^2 + y^2)^{1/2}dy = 0.$

4. **(a)** $\{6x/(3x^2 + y) + 4xy^3\}dx + \{1/(3x^2 + y) + 6x^2y^2 + 3y^2\}dy = 0;$
(b) $\{\sin(xy) + xy\cos(xy) + y^2\sin(xy)\}dx + \{x^2\cos(xy) + \cos(xy) - xy\sin(xy)\}dy = 0.$

5. **(a)** $\dfrac{3x^2}{2\sqrt{x^3 + y^2}}dx + \left\{\dfrac{y}{\sqrt{x^3 + y^2}} + 6y\right\}dy = 0;$
(b) $\{y/x + 2x\sinh(y^2)\}dx + \{\ln x + 2x^2y\cosh(y^2)\}dy = 0.$

6. **(a)** $\{4xy + 1/x\}dx + \{2x^2 - 1/y\}dy = 0;$
(b) $\{6xy - 2/(x^2y)\}dx + \{3x^2 - 2/(xy^2)\}dy = 0.$

7. **(a)** $\{2xy + 6/x\}dx + \{x^2 + 4/y\}dy = 0;$
(b) $\{2x/(2x + 3y^2) - 2x^2/(2x + 3y^2)^2 + 2\}dx - 6x^2y/(2x + 3y^2)^2dy = 0.$

8. **(a)** $\{(5/2)x^{3/2} + 14y^3\}dx + \{(3/2)\sqrt{y} + 42xy^2\}dy = 0;$
(b) $(y/x^2)\cos(y/x)dx + \{(1/x)\cos(y/x) + 6y\exp(y^2)\}dy = 0.$

5.7 Linear First Order Equations

The **standard form** of the **linear first order differential equation** is

standard form of linear first order equation

$$\frac{dy}{dx} + P(x)y = Q(x), \tag{25}$$

where $P(x)$ and $Q(x)$ are known functions. An **initial value problem** (**i.v.p**) for a linear first order ODE involves the specification of an initial condition

$$y(x_0) = y_0, \tag{26}$$

where this last condition means that $y = y_0$ when $x = x_0$. Thus, the solution of the initial value problem will evolve away from the point (x_0, y_0) in the (x, y)-plane as x increases from x_0.

To find the general solution of (25) we multiply the equation by a function $\mu(x)$, still to be determined, to obtain

$$\mu\frac{dy}{dx} + \mu P(x)y = \mu Q(x), \tag{27}$$

and seek a choice for μ that allows the left-hand side of (26) to be written as $d(\mu y)/dx$.

With this choice of μ, equation (27) becomes

$$\frac{d(\mu y)}{dx} = \mu Q(x), \tag{28}$$

so integrating with respect to x and dividing by μ shows the general solution of (25) to be

$$y(x) = \frac{C}{\mu(x)} + \frac{1}{\mu(x)}\int \mu(x)Q(x)dx, \tag{29}$$

where C is an arbitrary integration constant. Notice that it is essential to include the arbitrary integration constant *immediately* after the integration $\int \mu(x)Q(x)dx$ has been performed, and *before* dividing by $\mu(x)$; otherwise, the form of the general solution will be incorrect.

integrating factor

To make use of (29) it is necessary to determine the function $\mu(x)$ called the **integrating factor** for the linear first order ODE in (24). By definition

$$\frac{d(\mu y)}{dx} = \mu\frac{dy}{dx} + \mu P(x)y,$$

so after expanding the left-hand side this becomes

$$\mu\frac{dy}{dx} + y\frac{d\mu}{dx} = \mu\frac{dy}{dx} + \mu P(x)y.$$

Cancelling the terms $\mu dy/dx$ and dividing by y gives the following variables separable equation for the integrating factor $\mu(x)$:

$$\frac{d\mu}{dx} = \mu P(x).$$

This has the solution

$$\mu(x) = A\exp\left\{\int P(x)dx\right\},$$

where A is an arbitrary integration constant. As μ multiplies the entire equation (27), the choice of A is immaterial, so for simplicity we will always set $A = 1$ and take the **integrating factor** to be

finding the integrating factor

$$\mu(x) = \exp\left\{\int P(x)dx\right\}. \tag{30}$$

Inserting (30) into (29) shows the **general solution** of (25) to be

$$y(x) = C\exp\left\{-\int P(x)dx\right\} + \exp\left\{-\int P(x)dx\right\}\int \exp\left\{\int P(x)dx\right\} Q(x)dx. \tag{31}$$

If an initial value problem is involved in which the solution of (25) is required subject to the initial condition $y(x_0) = y_0$, the value of the arbitrary constant C in (31) must be chosen accordingly.

The form of the general solution in (31) is mainly of importance for theoretical reasons, because it shows that the general solution is the sum of a **complementary function**

complementary function, particular integral, and general solution

$$y_c(x) = C\exp\left\{-\int P(x)dx\right\} \tag{32}$$

that contains the arbitrary constant belonging to the general solution of (25), and a **particular integral**

$$y_p(x) = \exp\left\{-\int P(x)dx\right\}\int \exp\left\{\int P(x)dx\right\} Q(x)dx \tag{33}$$

that contains no arbitrary constant and is determined by the nonhomogeneous term $Q(x)$.

Substitution of $y_c(x)$ into the homogeneous form of (25) given by

$$\frac{dy}{dx} + P(x)y = 0$$

shows that $y_c(x)$ is its general solution. The general solution of the nonhomogeneous equation (25) is now seen to be the sum of the general solution of the homogeneous form of the equation, and a particular integral determined by the nonhomogeneous term. It will be shown later that this is the pattern of the general solution for all linear nonhomogeneous differential equations, no matter what their order.

Rather than trying to remember the form of general solution given in (31), it is better to obtain the solution by starting from the integrating factor $\mu(x)$ in (30) and integrating result (28), while not forgetting to include the arbitrary constant immediately after the integration before dividing by $\mu(x)$. For convenience, the steps in the determination of the general solution of (25) can be listed as follows.

steps used when solving a linear first order equation

Rule for solving linear first order equations

STEP 1 If the equation is not in standard form and is written

$$a(x)\frac{dy}{dx} + b(x)y = c(x),$$

divide by $a(x)$ to bring it to the standard form

$$\frac{dy}{dx} + P(x)y = Q(x),$$

with $P(x) = b(x)/a(x)$ and $Q(x) = c(x)/a(x)$

STEP 2 Find the integrating factor

$$\mu(x) = \exp\left\{\int P(x)dx\right\}.$$

STEP 3 Rewrite the original differential equation in the form

$$\frac{d(\mu y)}{dx} = \mu Q(x).$$

STEP 4 Integrate the equation in Step 3 to obtain

$$\mu(x)y(x) = \int \mu(x)Q(x)dx + C.$$

STEP 5 Divide the result of Step 4 by $\mu(x)$ to obtain the required general solution of the linear first order differential equation in Step 1.

STEP 6 If an initial condition $y(x_0) = y_0$ is given, the required solution of the i.v.p. is obtained by choosing the arbitrary constant C in the general solution found in Step 5 so that $y = y_0$ when $x = x_0$.

EXAMPLE 5.15

Solve the initial value problem

$$\cos x\frac{dy}{dx} + y = \sin x, \text{ subject to the initial condition } y(0) = 2.$$

Solution We follow the steps in the above rule.

STEP 1 When written in standard form the equation becomes

$$\frac{dy}{dx} + \frac{1}{\cos x}y = \tan x,$$

so $P(x) = 1/\cos x$ and $Q(x) = \tan x$.

STEP 2 The integrating factor

$$\mu(x) = \exp\left\{\int \frac{dx}{\cos x}\right\} = \exp\{\ln|\sec x + \tan x|\}$$
$$= \sec x + \tan x = \frac{1+\sin x}{\cos x}.$$

STEP 3 The original differential equation can now be written

$$\frac{d}{dx}\left[\left(\frac{1+\sin x}{\cos x}\right)y(x)\right] = \left(\frac{1+\sin x}{\cos x}\right)\tan x.$$

STEP 4 Integrating the result of Step 3 gives

$$\left(\frac{1+\sin x}{\cos x}\right)y(x) = \int \left(\frac{1+\sin x}{\cos x}\right)\tan x dx + C$$
$$= \int \sec x \tan x dx + \int \tan^2 x dx + C$$
$$= \sec x + \tan x - x + C = \frac{1+\sin x}{\cos x} - x + C.$$

STEP 5 Dividing the result of Step 4 by the integrating factor $\mu(x) = (1+\sin x)/\cos x$ shows that the required general solution is

$$y(x) = \frac{C\cos x}{1+\sin x} + 1 - \frac{x\cos x}{1+\sin x},$$

for x such that $1+\sin x \neq 0$.

The *complementary function* is seen to be

$$y_c(x) = \frac{C\cos x}{1+\sin x},$$

and the *particular integral* is

$$y_p(x) = 1 - \frac{x\cos x}{1+\sin x}.$$

STEP 6 The initial condition requires that $y = 2$ when $x = 0$, and the general solution is seen to satisfy this condition if $C = 1$, so the solution of the i.v.p. is

$$y(x) = 1 + \frac{(1-x)\cos x}{1+\sin x}.$$

■

EXAMPLE 5.16

An R–L circuit contains an inductor and resistor in series, and a current is made to flow through them by applying a voltage across the ends of the circuit. If the inductance varies linearly with time in such a way that $L(t) = L_0(1+kt)$, find the current $i(t)$ flowing in the circuit when $t > 0$, given that a constant voltage V_0 is applied at time $t = 0$ when $i(t) = 0$.

Solution The voltage change due to a current $i(t)$ flowing through the inductance is $d(L(t)i)/dt$, and from Ohm's law the corresponding voltage change across the resistance R is Ri, so as the sum of the these voltage changes must equal the imposed constant voltage V_0, the differential equation determining the current becomes

$$\frac{d}{dt}(L(t)i) + Ri = V_0 \quad \text{for} \quad t > 0.$$

Substituting for $L(t)$ and rearranging terms we arrive at the following linear first order variable coefficient nonhomogeneous equation for $i(t)$

$$\frac{di}{dt} + \left(\frac{kL_0 + R}{L_0(1+kt)}\right) i = \frac{V_0}{L_0(1+kt)},$$

subject to the initial condition $i(0) = 0$.

In the notation of this section $P(t) = \left(\dfrac{kL_0 + R}{L_0(1+kt)}\right)$ and $Q(t) = \dfrac{V_0}{L_0(1+kt)}$, so the integrating factor in Step 2 becomes

$$\mu(t) = \exp\left\{\int P(t)dt\right\} = (1+kt)^{[kL_0+R]/kL_0}.$$

Using $\mu(t)$ and $Q(t)$ in Step 4 and applying the initial condition $i(0) = 0$ then shows that the current $i(t)$ at a time $t > 0$ is determined by

$$i(t) = \left(\frac{V_0}{kL_0 + R}\right)\left(1 - (1+kt)^{\left(\frac{kL_0+R}{kL_0}\right)}\right). \qquad \blacksquare$$

Summary

The study of the linear first order differential equation considered in this section is important in its own right, and it also provides the key to understanding the nature of the solution of linear higher order differential equations. It was shown how, after an equation is written in standard form, it can be solved by means of an integrating factor that can be found directly from the coefficient of y in the equation.

EXERCISES 5.7

In Exercises 1 through 10 find the general solution for the linear first order differential equation, and check your result by using computer algebra.

1. $dy/dx + 2y = 1$.
2. $dy/dx + (1/x)y = x$.
3. $(x+1)dy/dx + y = 2x(x+1)$.
4. $x^2\,dy/dx + xy = x^2 \sin x$.
5. $x^2\,dy/dx - 2xy = 1 + x$.
6. $\sin x\,dy/dx - y\cos x = 2\sin^2 x$.
7. $x\,dy/dx + 2y = x^2$.
8. $(x+3)dy/dx - 2y = x + 3$.
9. $\sin x\,dy/dx - y = 2\sin x$.
10. $\sin x\,dy/dx + y = \sin x$.

In Exercises 11 through 16 solve the initial value problem for the linear first order differential equation, and check your result by using a computer algebra package.

11. $x\,dy/dx - y = x^2\cos x$, with $y(\pi/2) = \pi$.
12. $x^2\,dy/dx + 2xy = 2 + x$, with $y(1) = 1$.
13. $x\,dy/dx - 2y = 2 + x$, with $y(1) = 0$.
14. $x\,dy/dx + 2y = 2x^4$, with $y(1) = 1$.
15. $\sin x\,dy/dx + y\cos x = 2\sin^2 x$, with $y(\pi/2) = 0$.
16. $2\,dy/dx + y = x^2$, with $y(0) = 1$.
17. A 25-liter gas cylinder contains 80% oxygen and 20% helium. If helium is added at a rate of 0.2 liters a second, and the mixture is drawn off at the same rate, how long will it be before the cylinder contains 80% helium?
18. If in Exercise 17 the volume of the gas cylinder is 20 liters and initially it contains 90% oxygen and 10% helium, and the rate of supply of helium is q liters a second, what must be the value of q if the cylinder is to become 80% full of helium in 1 minute?

19. A particle of unit mass moves horizontally in a resisting medium with velocity $v(t)$ at time t with a resistance opposing the motion given by $kv(t)$, with $k > 0$. If the particle is also subject to an additional resisting force kt, write down the differential equation for $v(t)$, and hence find the value of k if the motion starts with $v(0) = v_0$, and at time $t = 1/k$ its velocity is $v(1/k) = \frac{1}{4}v_0$.

5.8 The Bernoulli Equation

The **Bernoulli equation** is a nonlinear first order differential equation with the standard form

standard form of the Bernoulli equation

$$\frac{dy}{dx} + P(x)y = Q(x)y^n, \quad (n \neq 1). \tag{34}$$

The substitution

$$u = y^{1-n} \tag{35}$$

reduces (34) to the linear first order ODE

$$\frac{1}{(1-n)}\frac{du}{dx} + P(x)u = Q(x), \tag{36}$$

and this can be solved by the method described in Section 5.7. Once the general solution $u(x)$ of (36) has been found, the general solution $y(x)$ of (34) follows by returning to the original dependent variable by making the substitution $u = y^{1-n}$.

When using the general solution in (36) it is important to write the Bernoulli equation in standard form before identifying $P(x)$, $Q(x)$, and n. However, if the form of the equation corresponding to (36) is derived directly, starting from the substitution $u = y^{1-n}$, there is no need for the equation to be in standard form.

The Bernoulli equation occurs in various applications of mathematics that involve some form of nonlinearity. It occurs, for example, in solid and fluid mechanics, where it is found to describe an important characteristic of special types of wave that propagate through space as time increases. To appreciate how this ODE enters into these problems, we consider a simple application to solid mechanics involving a long bar made of a composite material or a polymer whose properties are such that the extension caused by a force does not obey Hooke's law, and so is *not* proportional to the force. Materials of this type are said to be **nonlinearly elastic**. If such a bar receives a blow at one end a disturbance will propagate along it at a finite speed, so that at any instant of time there will be a region in the bar through which the disturbance has passed, and a region ahead of the disturbance through which it has still to pass. When the blow is not large, the propagating boundary between these two regions is called a **wavefront** and t the function representing the displacement at position x at any given time t will be continuous along the bar, though its derivative with respect to x will be discontinuous across the wavefront. The propagating jump in the derivative of the displacement with respect to x at the wavefront as a function of time is called an **acceleration wave**, and we will denote it

wavefront and acceleration wave

by $a(t)$. For many nonlinear materials the magnitude $a(t)$ of the acceleration wave obeys a Bernoulli equation of the form

$$\frac{da}{dt} + \mu(t)a = \beta(t)a^2. \tag{37}$$

It was shown by P. J. Chen (*Selected Topics in Wave Propagation*, Noordhoff, Leyden, 1976, p. 29) that $\mu(t)$ depends on the material properties of the medium through which the disturbance propagates and also the geometry involved, which in a one-dimensional case may be plane, cylindrically, or spherically symmetric, but that the function $\beta(t)$ depends only on the material properties of the medium. This same equation governs the behavior of acceleration waves in three space dimensions and time.

Because of the effects of nonlinearity, in many materials it is possible for the acceleration wave to strengthen as it propagates to the point at which the continuity of the displacement function breaks down and what is called a **shock wave** forms. When this occurs, the speed of propagation of disturbances and other physical quantities become discontinuous across the shock wave, and this in turn can lead to the fracture of the material. Once the material properties of such a medium are specified together with the nature of the initial disturbance, the Bernoulli equation in (37) can be used to determine whether or not a shock wave will form and, if it does, the point along the bar where this occurs.

EXAMPLE 5.17

examples of the Bernoulli equation

Solve the Bernoulli equation

$$\frac{da}{dt} + a = ta^2,$$

and find a condition that determines when the solution becomes unbounded.

Solution The equation is in standard form with $P(t) = 1$, $Q(t) = t$, and $n = 2$. Making the substitution $u = 1/a$ corresponding to (35) and substituting into (36) leads to the linear first order equation

$$\frac{du}{dt} - u = -t.$$

Solving this by the method described in Section 5.7 gives

$$u(t) = Ce^t + 1 + t,$$

so transforming back to the variable $a(t)$, we find that

$$a(t) = 1/(Ce^t + 1 + t).$$

The solution $a(t)$ of the Bernoulli equation will become unbounded at $t = t_c$ if t_c is a solution of the equation $C\exp(t_c) + 1 + t_c = 0$. This result shows that an acceleration wave starting at time $t = 0$ will decay instead of evolving into a shock wave if $C > 0$, because then the equation for t_c has no positive solution, whereas a shock wave will always form if $C < 0$.

Had $a(t)$ represented the magnitude of an acceleration wave, the development of an infinite gradient in the displacement corresponding to $a(t_c) = \infty$ would indicate shock formation. ■

EXAMPLE 5.18 Find the general solution of

$$\frac{dy}{dx} - 2y = xy^{1/2}.$$

Solution In terms of the standard form of the Bernoulli equation given in (34), $P(x) = -2$, $Q(x) = x$, and $n = 1/2$. However, rather than substituting into equation (36) to obtain a linear differential equation for $u(x)$, we will derive it directly starting from the substitution $u = y^{1/2}$, and differentiating it to find du/dx in terms of dy/dx. We have

$$\frac{du}{dx} = \frac{1}{2}y^{-1/2}\frac{dy}{dx} = \frac{1}{2u}\frac{dy}{dx}, \quad \text{so} \quad \frac{dy}{dx} = 2u\frac{du}{dx}.$$

Substituting for y and dy/dx in the Bernoulli equation and cancelling a factor $2u$ gives the following linear equation (compare it with (36) after substituting for $P(x)$, $Q(x)$ and n):

$$\frac{du}{dx} - u = \frac{1}{2}x.$$

The method of Section 5.6 shows this equation to have the general solution

$$u(x) = Ce^x - (1/2)(1 + x),$$

so as $u = y^{1/2}$, the required general solution of the Bernoulli equation is

$$y(x) = [Ce^x - (1/2)(1 + x)]^2. \quad \blacksquare$$

JACOB BERNOULLI (1654–1705)
A Swiss mathematician born in Basel where he was professor of mathematics until his death. He was a member of one of the most distinguished families of mathematicians in all of the history of mathematics. His most important contributions were to the theory of probability and the calculus and theory of elasticity. Other members of the family contributed to many different parts of mathematics including hydrodynamics and the calculus of variations.

Summary

In a sense, the Bernoulli equation, which is a nonlinear first order differential equation, stands on the boundary between linear and nonlinear first order differential equations, so for this and other reasons it is important in applications. It arises in different applications, many of which themselves arise from problems bordering on linear and nonlinear regimes. This section showed how a straightforward change of variable transforms a Bernoulli equation into a linear first order differential equation that can then be solved by the method of Section 5.6.

EXERCISES 5.8

In Exercises 1 through 8 find the general solution of the Bernoulli equation.

1. $dy/dx + 2y = 2xy^{1/2}$.
2. $dy/dx + y = 3y^2$.
3. $dy/dx - y = 2xy^{3/2}$.
4. $x\,dy/dx + y = xy^2$.
5. $dy/dx + 2y\sin x = 2y^2\sin x$.
6. $x\,dy/dx + y = 2xy^{1/2}$.
7. $x\,dy/dx - 2y = xy^{3/2}$.
8. $dy/dx + 4xy = xy^3$.

9. A model for the variation of a finite amount of stock $n(t)$ in a warehouse as a function of the time t caused by the supply of fresh stock and its removal by demand is

$$\frac{dn}{dt} = (a - bn)n \quad \text{with the constants } a, b > 0,$$

where $n(0) = n_0$. Find $n(t)$ and discuss the nature of the change in the stock level as a function of time according as n_0 is less than a/b, equal to a/b, or greater than a/b.

10.* This exercise concerns water in a canal of variable depth with the x-axis taken along the canal in the equilibrium surface of the water, and the y-axis vertically downwards. Let the equilibrium depth of water in a channel be $h(x)$, and the cross-sectional area of water in the canal be a slowly varying function $W(x)$. When a water wave advances along the channel into water at rest there will be a change of acceleration across the advancing line (wavefront) that separates the disturbed water from the undisturbed water. Such an advancing disturbance is called an **acceleration wave**. If the change in acceleration across the wavefront at point x along the channel is $a(x)$, it can be shown that the strength $a(x)$ of the acceleration wave obeys the Bernoulli equation

$$\frac{da}{dx} + \left(\frac{3h'}{4h} + \frac{W'}{2W}\right)a + \frac{3a^2}{2h} = 0.$$

If the initial condition for $a(x)$ is $a(0) = a_0$, then a wave of **elevation** wave is one for which $a_0 < 0$, and a wave of **depression** is one for which $a_0 > 0$. In this approximation the wave will **break**, due to the water surface becoming vertical at the wavefront if, after propagating a critical distance x_c along the channel, the strength of the acceleration $a(x_c) = \infty$.

(i) Find $a(x)$ in terms of $a_0 = a(0)$, $h_0 = h(0)$ and $W_0 = W(0)$.

(ii) Discuss the breaking and non-breaking of waves of elevation and depression.

(iii) If the water shelves to zero at $x = l$, so that $h(l) = 0$, find a condition that ensures the wave breaks before $x = l$.

5.9 The Riccati Equation

The **Riccati equation** is an important nonlinear equation with the standard form

standard form of the Riccati equation

$$\frac{dy}{dx} + P(x)y + R(x)y^2 = Q(x). \tag{38}$$

Its significance derives from the fact that it stands at the boundary between linear and nonlinear equations, and it occurs in various applications of mathematics that involve nonlinear problems. The Riccati equation reduces to a linear first order equation when $R(x) \equiv 0$, and to a Bernoulli equation when $Q(x) \equiv 0$.

Obtaining the general solution of a Riccati equation is difficult, but the task is simplified if a particular solution is known, or can be found by inspection. If a particular solution is $y_1(x)$ is known, then

(i) The substitution $y = y_1 + 1/u$ reduces the equation to a linear first order equation.

substitutions that simplify the Riccati equation

(ii) The substitution $y = y_1 + u$ reduces the equation to a Bernoulli equation.

(iii) The general substitution

$$y = \frac{1}{R(x)z}\frac{dz}{dx}$$

reduces the Riccati equation to the linear homogeneous second order ODE

$$\frac{d^2z}{dx^2} + \left\{ P(x) - \frac{R'(x)}{R(x)} \right\} \frac{dz}{dx} - R(x)Q(x)z = 0$$

discussed in Chapters 6 and 8.

Substitution (i) is often the most convenient one to use, as will be seen from the next example.

EXAMPLE 5.19 Find the general solution of the Riccati equation

$$\frac{dy}{dx} + x^2y - xy^2 = 1.$$

Solution Inspection shows that $y_1(x) = x$ is a particular solution, so we make the substitution $y = x + 1/u$, from which it follows that

$$\frac{dy}{dx} = 1 - \frac{1}{u^2}\frac{du}{dx},$$

and after substitution for y and dy/dx in the Riccati equation it reduces to the linear ODE

$$\frac{du}{dx} + x^2u = -x.$$

Solving this by the method of Section 5.6 gives

$$u(x) = C\exp(-x^3/3) - \exp(-x^3/3)\int x\exp(x^3/3)dx,$$

where the integral in the last term cannot be expressed in terms of elementary functions. Transforming back to the variable $y(x)$ shows the general solution of the Riccati equation to be

$$y(x) = x + \frac{\exp(x^3/3)}{C - \int x\exp(x^3/3)dx}.$$

It is not unusual for solutions of ODEs to give rise to functions such as $\int x\exp(x^3/3)dx$ that have no representation in terms of known functions, because not all functions have antiderivatives that are expressible in terms of elementary functions. ■

JACOPO FRANCESCO (COUNT) RICCATI (1676–1754)
An Italian mathematician whose main contributions to mathematics were in the field of differential equations, though he also contributed to geometry and the study of acoustics.

Additional information relevant to the material in Sections 5.4 to 5.9 is to be found in the appropriate chapters of any one of references [3.3] to [3.5], [3.15], [3.16], and [3.19]. A sophisticated and extremely enlightening discussion of ordinary differential equations is to be found in reference [3.1] that considers not only first order equations, but also higher order equations and systems.

Summary

This section introduced the Riccati equation, of which the Bernoulli equation is a special case. Solving the Riccati equation is difficult, but some substitutions were given that simplify this task when one solution of the Riccati equation is already known, possibly by inspection.

EXERCISES 5.9

1. Show that the substitution $y = y_1 + 1/u$ reduces the Riccati equation in (38) to a linear first order equation.

2. Show that the substitution $y = y_1 + u$ reduces the Riccati equation in (38) to a Bernoulli equation.

In Exercises 3 through 6 verify that $y_1(x)$ is a solution of the Riccati equation and use it to find the general solution of the equation.

3. $dy/dx + 2x^2y - 2xy^2 = 1,$ with $y_1(x) = x.$

4. $dy/dx + 2y^2 - y = 1,$ with $y_1(x) \equiv 1.$

5. $dy/dx - 2y^2 + 3y = 1,$ with $y_1(x) \equiv 1.$

6. $dy/dx - 3x^2y + 3xy^2 = 1,$ with $y_1(x) = x.$

7. Verify that the substitution

$$y = \frac{1}{R(x)z}\frac{dz}{dx}$$

reduces the Riccati equation (38) to the linear homogeneous second order ODE

$$\frac{d^2z}{dx^2} + \left\{P(x) - \frac{R'(x)}{R(x)}\right\}\frac{dz}{dx} - R(x)Q(x)z = 0.$$

5.10 Existence and Uniqueness of Solutions

existence and uniqueness

The questions of whether a solution to an initial value problem for a first order differential equation can be found and, when a solution does exist, whether it is the only solution are of fundamental importance in the theory of differential equations, and also in their applications. Establishing that a solution to an initial value problem can be found is called the **existence** problem, while ensuring that when a solution exists it is the only one is called the **uniqueness** problem. To show that the questions of existence and uniqueness arise even with very simple initial value problems we examine the following two examples.

Let us consider the initial value problem

$$\frac{dy}{dx} = \frac{4}{3}y^{1/4}, \quad \text{with } y(0) = -1,$$

involving a variables separable equation. Integration shows the general solution to be

$$y^3 = (x + C)^4,$$

from which it can be seen that y is essentially nonnegative. Clearly there can be no solution to this equation such that $y = -1$ when $x = 0$, so this is an example of an initial value problem that has *no* solution. Had the initial condition been $y(0) = 1$ the unique solution would have been

$$y^3 = (x + 1)^4.$$

In fact this equation has a solution for any initial condition in which $y(x)$ is *positive*, but no solution when it is *negative*. This is hardly surprising, because had we examined the function $y^{1/4}$ carefully before proceeding with the integration we would have seen that it is a complex number whenever y is negative. Sometimes,

as here, an inspection of the initial condition and the equation can show in advance whether or not the condition is appropriate, but more frequently constraints on an initial condition that allow a solution to the differential equation to exist only emerge when the form of the solution is known.

To illustrate nonuniqueness, we need only consider the differential equation

$$\frac{dy}{dx} = 3y^{2/3}, \text{ subject to the initial condition } y(0) = 0.$$

The equation is variables separable, and integration shows it has the solution $y = x^3$, but this is not the only solution because it also has the singular, though somewhat uninteresting, solution $y = 0$.

However, these are not the only two solutions, because for any $a > 0$ the function

$$y(x) = \begin{cases} 0, & x < a \\ (x-a)^3, & x \geq a \end{cases}$$

is continuous, has a continuous first derivative, and satisfies both the differential equation and the initial condition, showing that it also is a solution. As $a > 0$ is arbitrary, we see that $y(x)$ is a one-parameter family of solutions, so clearly this initial value problem does not have a unique solution.

The following theorem on existence and uniqueness is stated without proof (see, for example, references [3.1],[3.3],[3.4],[3.10] and [3.12]). It is important to appreciate that though the conditions in the theorem are *sufficient* to ensure existence and uniqueness, they are not *necessary* conditions, as examples can be constructed that fail to satisfy the conditions of the theorem, but nevertheless have a unique solution.

THEOREM 5.2

conditions that definitely ensure existence and uniqueness

Existence and uniqueness of solutions Let $f(x, y)$ be a continuous and bounded function of x and y in a rectangular region R of the (x, y)-plane that contains a given point (x_0, y_0). Then for some suitably small positive number h the initial value problem

$$\frac{dy}{dx} = f(x, y), \quad \text{with} \quad y(x_0) = y_0$$

has at least one solution within the open interval $x_0 - h < x < x_0 + h$. If, in addition, $\partial f/\partial y$ is continuous and bounded in R, the solution is unique in an open interval centered on x_0 that may lie within the interval $x_0 - h < x < x_0 + h$. ■

Let us apply this theorem to the initial value problem

$$\frac{dy}{dx} = 3y^{2/3}, \quad \text{with } y(0) = 0,$$

that we have just shown does not have a unique solution. The function $f(x, y) = 3y^{2/3}$ is continuous in any neighborhood of the origin where the initial condition is given, but $\partial f/\partial y = 2y^{-1/3}$ is unbounded at the origin. So the first condition of Theorem 5.2 is satisfied but the second is not, showing that although this initial value problem has a solution, it is not unique.

Summary

This section described what is meant by the existence of a solution of a differential equation, and the uniqueness of a solution that is usually expected in applications to physical problems. A theorem, stated without proof, was given that guarantees both the existence and uniqueness of a solution. However, the conditions of the theorem are more restrictive than necessary, so equations can be found that while not satisfying the conditions of the theorem nevertheless have a solution, and it is unique.

EXERCISES 5.10

In Exercises 1 through 6, find any points at which the imposition of initial conditions will not lead to a unique solution.

1. $dy/dx = (1 - x)^{1/2}$.

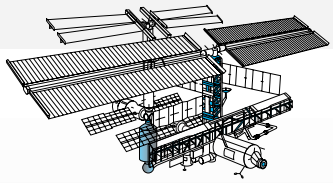
2. $dy/dx = xy + 1$.

3. $dy/dx = x^2 + y^2$.

4. $dy/dx = (x^2 + y^2 - 1)^{-1/2}$.

5. $dy/dx = -y/x$.

6. $dy/dx = x \ln|1 - y^2|$.



CHAPTER 5
TECHNOLOGY PROJECTS

Project 1

Solution of First Order Linear Differential Equation

The purpose of this project is to use computer algebra to solve a first order equation step by step from first principles, and then to obtain the same result by means of a computer software ODE solver.

1. Given the linear first order differential equation

$$y' + (3x^2 \sin x)y = 2x^2 \sin x,$$

use computer integration to find the general solution by reproducing the steps in the rule for the solution by means of an integrating factor given in Section 5.6, and check the result by substitution into the differential equation.

2. Use a computer ODE solver to find the general solution and confirm that it is the same as the result obtained in step 1.

Project 2

Direction Fields and Integral Curves

The purpose of the following project is to gain insight into the relationship between direction fields and integral curves by using a computer package to plot the direction fields for two nonlinear first order differential equations, and then to add to the direction field plots some typical integral curves obtained by using a standard numerical ODE solver package.

1. Construct the direction field for the nonlinear ODE

$$y' = \sin\left(\frac{1}{2}x\right)\cos\left(\frac{1}{2}x + y\right) \quad \text{for} \quad -6 \le x \le 6, \quad -6 \le y \le 6.$$

2. Use a standard ODE numerical solver package to find the solutions (the integral curves) through the points $(-6, -4)$, $(-6, -2)$, $(-6, 2)$, $(-6, 4)$. Superimpose the integral curves on the direction field and compare them with the arrows in the direction field.

3. Repeat Steps 1 and 2, but this time using the nonlinear ODE

$$y' = x \sin(y-1)/(3 + \cos x) \quad \text{for} \quad -6 \le x \le 6, \quad -6 \le y \le 6.$$

Project 3

Direction Fields and Isoclines

An *isocline* is a curve in the direction field of the differential equation $y' = f(x, y)$ at each point of which the slope of the direction field has the same constant value. This means that wherever a solution curve of the equation intersects an isocline, its tangent will have the same slope. The isoclines of the differential equation $y' = f(x, y)$ are the curves $k = f(x, y)$, where k is the slope (gradient) of all solution curves at the points where they intersect the isocline. In general an isocline is not a solution curve and, depending on the function $f(x, y)$, there may be no isoclines for some values of the constant k. The purpose of this project is to construct the direction field for an ODE, and to superimpose on it some representative isoclines and solution curves to illustrate their interrelationship.

1. Use computer algebra to construct the direction field for the ordinary differential equation

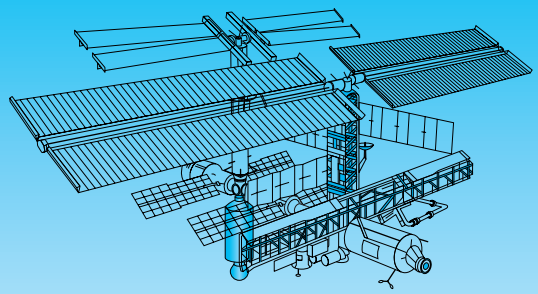
$$y' = x^2 - y - 1 \quad \text{for} \quad -2 \le x \le 2, -2 \le y \le 2,$$

and superimpose on the direction field the isoclines corresponding to $k = -1, 0, 1, 2$. Verify that all arrows intersecting an isocline are parallel.

2. Use a standard ODE numerical solver package to find the solutions through the points $(-2, -1.5)$, $(-2, -0.5)$, $(-2, 0.5)$, $(-2, 1.5)$. Superimpose the solution curves on the isoclines found in Step 1 and confirm that the tangents to solution curves where they intersect an isocline are all parallel.

This Page Intentionally Left Blank

CHAPTER 6



Second and Higher Order Linear Differential Equations and Systems

Linear second order differential equations with constant coefficients are the simplest of the higher order differential equations, and they have many applications. They are of the general form $y'' + Ay' + By = F(x)$ with A and B constants and $F(x)$, called the nonhomogeneous term, a known function of x. The equation is called nonhomogeneous when $F(x)$ is not identically zero; otherwise, it is called homogeneous. All general solutions are shown to be the sum of two quite different parts, one being a solution of the homogeneous equation called the complementary function that contains the expected two arbitrary constants of integration, and the other a special solution called a particular integral that depends only on $F(x)$ and contains no arbitrary constants.

Methods are developed for the solution of homogeneous and nonhomogeneous second order equations and for the solution of associated initial value problems. Particular attention is paid to the second order equations that describe oscillatory phenomena, because equations of this type arise in practical problems involving oscillations in electrical circuits, in the description of many types of mechanical vibration, and elsewhere. It is shown that in stable oscillatory motions the particular integral describes the start-up of an initial value problem, after which it decays, leaving only the complementary function that describes the long-term behavior known as the steady state solution.

The methods of solution for second order equations developed in this chapter include the simplest one, called the method of undetermined coefficients; the powerful method of variation of parameters; and a related method involving a function called the Green's function that is independent of the nonhomogeneous term $F(x)$.

Various useful special cases of second order equations are considered, after which higher order linear differential equations and first order systems are introduced and solved, the solutions of which have the same general structure as the second order equations. Matrix methods are introduced for the description and solution of first order systems of equations. The chapter concludes with a discussion of linear autonomous systems of equations, followed by a brief introduction to nonlinear autonomous systems that arise in many practical problems and can lead to oscillatory solutions of a nonlinear nature. The general behavior of solutions of both types of autonomous system is described in an interesting and useful geometrical manner involving what are called trajectories in the phase plane.

6.1 Homogeneous Linear Constant Coefficient Second Order Equations

linear constant coefficient second order equation

The simplest general higher order homogeneous differential equation that occurs in applications is the **linear constant coefficient second order equation**

$$\frac{d^2y}{dx^2} + A\frac{dy}{dx} + By = 0. \tag{1}$$

Equations like this were derived in Section 5.2(d), where they were shown to describe the motion of a mass–spring system subject to frictional resistance, and also the variation of charge in an R–L–C electric circuit. The equation also describes the pendulum-like motion of a load suspended from a crane that is set in motion when the crane rotates to a new position and soon stops. The motion can be modeled as shown in Fig. 6.1, where ℓ is the length of the crane cable, m is the load, F is the resisting frictional force exerted by the air due to motion, and θ is the angular deflection of the cable from the vertical.

The angular momentum of the load about a line through the support point of the cable at O normal to the plane of motion is $m\ell^2(d\theta/dt)$, so the rate of change of angular momentum about O is $m\ell^2(d^2\theta/dt^2)$. The moments acting to restore the load to its equilibrium position at Q are due to the air resistance F opposing the motion and the turning moment of the gravitational force mg about O. If the air resistance acting on the load is proportional to the speed of the load, and the constant of proportionality is μ, the resisting frictional force is $F = \mu\ell(d\theta/dt)$, so the restoring moment exerted by F about O is $\ell F = \mu\ell^2(d\theta/dt)$. The turning moment exerted by the gravitational force mg about O is $mg\ell\sin\theta$, so equating the rate of change of angular momentum to the sum of the two restoring moments gives

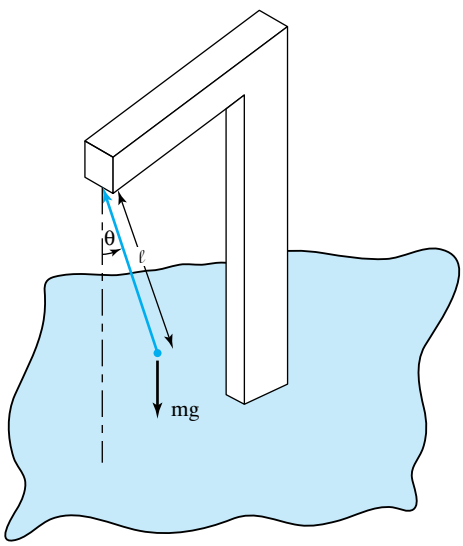


FIGURE 6.1 A deflected load supported by a crane cable.

the equation of motion

$$m\ell^2\frac{d^2\theta}{dt^2} = -\mu\ell^2\frac{d\theta}{dt} - mg\ell\sin\theta.$$

The negative signs on the right are necessary because the restoring moments act in the opposite sense to that of the rate of change of angular momentum.

When the angle of swing is small $\sin\theta$ can be approximated by θ, and the equation of motion simplifies to

$$\frac{d^2\theta}{dt^2} + \frac{\mu}{m}\frac{d\theta}{dt} + \frac{g}{\ell}\theta = 0.$$

Because of its many applications we start our discussion of higher order equations by examining the properties and general solution of equation (1).

Let $y_1(x)$ and $y_2(x)$ be any two solutions of (1). Then because each function satisfies the differential equation, it follows that

$$\frac{d^2y_1}{dx^2} + A\frac{dy_1}{dx} + By_1 = 0 \quad \text{and} \quad \frac{d^2y_2}{dx^2} + A\frac{dy_2}{dx} + By_2 = 0. \tag{2}$$

Now consider the linear combination of the two solutions

$$y(x) = c_1y_1(x) + c_2y_2(x), \tag{3}$$

where c_1 and c_2 are arbitrary constants. Substituting (3) into (1) and grouping terms gives

$$\frac{d^2[c_1y_1 + c_2y_2]}{dx^2} + A\frac{d[c_1y_1 + c_2y_2]}{dx} + B[c_1y_1 + c_2y_2]$$
$$= c_1\left[\frac{d^2y_1}{dx^2} + A\frac{dy_1}{dx} + By_1\right] + c_2\left[\frac{d^2y_2}{dx^2} + A\frac{dy_2}{dx} + By_2\right] = 0,$$

because each of the bracketed groups of terms vanishes on account of (2). This has shown that $y(x) = c_1y_1(x) + c_2y_2(x)$ is also a solution of (1).

linear superposition, dependence, and independence

This last result is described by saying equation (1) allows the **linear superposition** of solutions and it means that the sum of solutions is again a solution. Later we will see that linear superposition of solutions is a fundamental property of all homogeneous linear equations, including those with variable coefficients.

Two functions $y_1(x)$ and $y_2(x)$ are said to be **linearly independent** over an interval $a \le x \le b$ if the equation

$$c_1y_1(x) + c_2y_2(x) = 0 \tag{4}$$

is only true for all x in the interval if $c_1 = c_2 = 0$. The functions are said to be **linearly dependent** if (4) is true for some nonvanishing constants c_1 and c_2.

When the functions are linearly dependent, provided $c_1 \neq 0$, equation (4) can be written

$$y_1(x) = -\frac{c_2}{c_1}y_2(x),$$

with a corresponding result

$$y_2(x) = -\frac{c_1}{c_2} y_1(x),$$

if $c_2 \neq 0$, showing that in each case the linear dependence of the functions means they are proportional. We have established the following simple test.

simple test for linear independence

Test for linear independence of $y_1(x)$ and $y_2(x)$ over $a \leq x \leq b$

The two functions $y_1(x)$ and $y_2(x)$ will be linearly independent over $a \leq x \leq b$ if they are not proportional over the interval; otherwise, they will be linearly dependent.

EXAMPLE 6.1

Apply the test for linear independence to the following pairs of functions.

(a) e^x and e^{2x} are linearly independent for all x because $e^{2x}/e^x = e^x$ is defined for all x and e^x is *not* a constant.

(b) $\ln x^2$ and $\ln x^3$ are linearly dependent for $x > 0$, because $\ln x^2 = 2 \ln x$ and $\ln x^3 = 3 \ln x$, so $\ln x^2 / \ln x^3 = 2/3$ is a constant, and the logarithmic function is defined for $x > 0$.

(c) $\sinh 2x$ and $\sinh x \cosh x$ are linearly dependent for all x because $\sinh 2x = 2 \sinh x \cosh x$. ■

The notion of the linear independence of functions is of special significance when the functions are solutions of homogeneous differential equations. This is because it will be seen later that all particular solutions of such differential equations can be represented in the form of suitable linear combinations of as many linearly independent solutions as the equation allows. In fact, the number of linearly independent solutions is equal to the order of the differential equation, so the second order differential equation (1) has *two* linearly independent solutions. So, if $y_1(x)$ and $y_2(x)$ are linearly independent solutions of (1), and c_1 and c_2 are arbitrary constants, the **general solution** of (1) from which all particular solutions can be obtained can be written

general solution

$$y(x) = c_1 y_1(x) + c_2 y_2(x). \tag{5}$$

The justification of this assertion will be postponed until the nature of the linearly independent solutions of (1) has been established.

EXAMPLE 6.2

Direct substitution of the functions $y_1(x) = \sin 2x$ and $y_2(x) = \cos 2x$ into the second order differential equation

$$y'' + 4y = 0$$

confirms that they are solutions. The functions are linearly independent for all x because they are not proportional, so

$$y(x) = c_1 \cos 2x + c_2 \sin 2x$$

is the general solution of the differential equation. ■

We will now find the general solution of (1), and when doing so use will be made of the fact that if $y(x) = ce^{\lambda x}$, with c and λ constants, then

$$\frac{dy}{dx} = \frac{d[ce^{\lambda x}]}{dx} = c\lambda e^{\lambda x} \quad \text{and} \quad \frac{d^2y}{dx^2} = \frac{d^2[ce^{\lambda x}]}{dx^2} = c\lambda^2 e^{\lambda x}.$$

Substituting these results into (1) leads to the equation

$$(\lambda^2 + A\lambda + B)e^{\lambda x} = 0.$$

However, the factor $e^{\lambda x}$ is nonvanishing for all x, so after its cancellation this equation is seen to be equivalent to the quadratic equation for λ

$$\lambda^2 + A\lambda + B = 0. \tag{6}$$

When the quadratic equation (6) has two distinct (different) roots λ_1 and λ_2, the functions $y_1(x) = \exp(\lambda_1 x)$ and $y_2(x) = \exp(\lambda_2 x)$ will be linearly independent for all x, because $y_1(x)/y_2(x) = \exp[(\lambda_1 - \lambda_2)x]$ is not constant. Thus, then $\exp(\lambda_1 x)$ and $\exp(\lambda_2 x)$ are linearly independent solutions of (1), so the general solution is

$$y(x) = c_1 \exp(\lambda_1 x) + c_2 \exp(\lambda_2 x), \tag{7}$$

where c_1 and c_2 are arbitrary constants.

It is now necessary to introduce the type of initial conditions that are appropriate for (1). As (1) is a second order differential equation, it relates $y(x)$, $y'(x)$, and $y''(x)$, so it follows that suitable initial conditions will be the specification of $y(x)$ and $y'(x)$ at some point $x = a$. Then the value of $y''(a)$ cannot be assigned arbitrarily, because the differential equation itself will determine its value in terms of $y(a)$ and $y'(a)$. The solution of (1) satisfying these initial conditions can be found from the general solution (7) by determining c_1 and c_2 from the two equations:

initial conditions

Initial condition on y(x)

$$y(a) = c_1 \exp(\lambda_1 a) + c_2 \exp(\lambda_2 a),$$

Initial condition on y′(x)

$$y'(a) = \lambda_1 c_1 \exp(\lambda_1 a) + \lambda_2 c_2 \exp(\lambda_2 a). \tag{8}$$

When we considered systems of linear algebraic equations in Chapter 3, it was shown that equations (8) will determine c_1 and c_2 uniquely if the determinant of the coefficients of c_1 and c_2 is nonvanishing. Thus, the specification of $y(a)$ and $y'(a)$ will be appropriate as initial conditions if

$$\Delta = \begin{vmatrix} \exp(\lambda_1 a) & \exp(\lambda_2 a) \\ \lambda_1 \exp(\lambda_1 a) & \lambda_2 \exp(\lambda_2 a) \end{vmatrix} \neq 0. \tag{9}$$

Expanding the determinant gives $\Delta = (\lambda_2 - \lambda_1)\exp[(\lambda_1 + \lambda_2)a]$. However, by hypothesis $\lambda_1 \neq \lambda_2$, while $\exp[(\lambda_1 + \lambda_2)a]$ never vanishes, so $\Delta \neq 0$. The particular solution satisfying the initial conditions follows by using the values of c_1 and c_2 found from (8) in the general solution (7).

EXAMPLE 6.3 Find the solution of the initial value problem

$$y'' + 4y = 0, \quad \text{if} \quad y(\pi/4) = 1 \quad \text{and} \quad y'(\pi/4) = 1.$$

Solution In Example 6.2 direct substitution has already been used to show that $\cos 2x$ and $\sin 2x$ are linearly independent solutions of the differential equation, so its general solution is

$$y(x) = c_1 \cos 2x + c_2 \sin 2x,$$

from which it follows by differentiation that

$$y'(x) = -2c_1 \sin 2x + 2c_2 \cos 2x.$$

Imposing the initial condition on $y(x)$ at $x = \pi/4$ leads to the following equation that must be satisfied by c_1 and c_2:

$$1 = c_1 \cos \pi/2 + c_2 \sin \pi/2.$$

Similarly, imposing the initial condition on $y'(x)$ at $x = \pi/4$ leads to the second condition that must be satisfied by c_1 and c_2:

$$1 = -2c_1 \sin \pi/2 + 2c_2 \cos \pi/2.$$

These equations have the solution $c_1 = -1/2$ and $c_2 = 1$, so the particular solution satisfying the initial conditions $y(\pi/4) = 1$ and $y'(\pi/4) = 1$ is

$$y(x) = \sin 2x - \frac{1}{2}\cos 2x. \qquad \blacksquare$$

The quadratic equation determining the permissible values of λ in the exponential solutions $y_1(x) = \exp(\lambda_1 x)$ and $y_2(x) = \exp(\lambda_2 x)$ of differential equation (1), namely,

$$\lambda^2 + A\lambda + B = 0, \tag{10}$$

characteristic equation

is called the **characteristic equation** of the differential equation. Its two roots,

$$\lambda_1 = \frac{-A + \sqrt{A^2 - 4B}}{2} \quad \text{and} \quad \lambda_2 = \frac{-A - \sqrt{A^2 - 4B}}{2}, \tag{11}$$

are the values of λ to be used in the general solution (7). When the roots λ_1 and λ_2 are real and distinct, the functions

$$y_1(x) = \exp(\lambda_1 x) \quad \text{and} \quad y_2(x) = \exp(\lambda_2 x) \tag{12}$$

are said to form a **basis** for the solution space of (1). This means that the solution of every initial value problem for (1) can be obtained from the linear combination $y(x) = c_1 \exp(\lambda_1 x) + c_2 \exp(\lambda_2 x)$ by assigning suitable values to c_1 and c_2.

A comparison of differential equation (1) and its characteristic equation (10) shows the characteristic equation can be written down immediately from the differential equation by simply replacing y by 1, dy/dx by λ and d^2y/dx^2 by λ^2. It is

usual to use this method when obtaining the characteristic equation, as it avoids the unnecessary intermediate steps involved when substituting $y(x) = \exp(\lambda x)$.

Three different cases must now be considered, according to whether (i) λ_1 and λ_2 are real and distinct ($\lambda_1 \neq \lambda_2$), (ii) λ_1 and λ_2 are complex conjugates, or (iii) the possibility, excluded so far, that λ_1 and λ_2 are real and equal, so $\lambda_1 = \lambda_2 = \mu$, say.

Case (I) (Real and Distinct Roots)

how a solution depends on the roots

This case corresponds to the condition $A^2 - 4B > 0$, with

$$\lambda_1 = \frac{-A + \sqrt{A^2 - 4B}}{2} \quad \text{and} \quad \lambda_2 = \frac{-A - \sqrt{A^2 - 4B}}{2}. \tag{13}$$

No more need be said about this case because it has already been established that the functions $\exp(\lambda_1 x)$ and $\exp(\lambda_2 x)$ form a basis for the solution space of (1), which thus has the general solution

$$y(x) = c_1 \exp(\lambda_1 x) + c_2 \exp(\lambda_2 x).$$

Case (II) (Complex Conjugate Roots)

This case corresponds to the condition $A^2 - 4B < 0$. A real solution $y(x)$ corresponding to complex conjugate roots λ_1 and λ_2 is only possible if the arbitrary constants c_1 and c_2 are themselves complex conjugates. A routine calculation shows that if $\lambda_1 = \alpha + \mathrm{i}\beta$ and $\lambda_2 = \alpha - \mathrm{i}\beta$, with

$$\alpha = -(1/2)A, \qquad \beta = (1/2)(4B - A^2)^{1/2}, \tag{14}$$

the two corresponding linearly independent solutions are

$$y_1(x) = e^{\alpha x} \cos \beta x \quad \text{and} \quad y_2(x) = e^{\alpha x} \sin \beta x. \tag{15}$$

A basis for the solution space of (1) is formed by the functions $e^{\alpha x} \cos \beta x$ and $e^{\alpha x} \sin \beta x$, corresponding to a general solution of the form

$$y_1(x) = e^{\alpha x}[c_1 \cos \beta x + c_2 \sin \beta x]. \tag{16}$$

The calculation required to establish the form of this result is left as an exercise.

Case (III) (Equal Real Roots)

This case corresponds to the condition $A^2 - 4B = 0$, with

$$\mu = \lambda_1 = \lambda_2 = -(1/2)A. \tag{17}$$

In this case only the one exponential solution

$$y_1(x) = e^{\mu x} \tag{18}$$

can be found.

However, substitution of the function

$$y_2(x) = xe^{\mu x} \tag{19}$$

into the differential equation shows that it is also a solution. The functions $y_1(x)$ and $y_2(x)$ are linearly independent because $y_2(x)/y_1(x) = x$ is not a constant, so in this case a basis for the solution space of (1) is formed by the functions $e^{\mu x}$ and $xe^{\mu x}$, with the corresponding general solution

$$y(x) = (c_1 + c_2 x)e^{\mu x}. \tag{20}$$

Summary of the forms of solution of $y'' + Ay' + By = 0$

summary of types of solution

Characteristic equation: $\lambda^2 + A\lambda + B = 0$

Case (I) $A^2 - 4B > 0$. The general solution is

$$y(x) = c_1 \exp(\lambda_1 x) + c_2 \exp(\lambda_2 x), \quad \text{with}$$
$$\lambda_1 = \frac{-A + \sqrt{A^2 - 4B}}{2} \quad \text{and} \quad \lambda_2 = \frac{-A - \sqrt{A^2 - 4B}}{2}.$$

Case (II) $A^2 - 4B < 0$. The general solution is

$$y_1(x) = e^{\alpha x}[c_1 \cos \beta x + c_2 \sin \beta x], \quad \text{with}$$
$$\alpha = -(1/2)A \quad \text{and} \quad \beta = (1/2)(4B - A^2)^{1/2}.$$

Case (III) $A^2 = 4B$. The general solution is

$$y(x) = (c_1 + c_2 x)e^{\mu x}, \quad \text{with} \quad \mu = -(1/2)A.$$

EXAMPLE 6.4

Find the general solution and hence solve the stated initial value problem for

(i) $y'' + y' - 2y = 0$, with $y(0) = 1$ and $y'(0) = 2$;
(ii) $y'' + 2y' + 4y = 0$, with $y(0) = 2$ and $y'(0) = 1$;
(iii) $y'' + 4y' + 4y = 0$, with $y(0) = 3$ and $y'(0) = 1$.

Solution

(i) The characteristic equation is

$$\lambda^2 + \lambda - 2 = 0,$$

with the roots $\lambda_1 = 1, \lambda_2 = -2$, so this is Case (I). The general solution is

$$y(x) = c_1 e^x + c_2 e^{-2x}.$$

The initial condition $y(0) = 1$ is satisfied if

$$1 = c_1 + c_2,$$

while the initial condition $y'(0) = 2$ is satisfied if

$$2 = c_1 - 2c_2.$$

These equations have the solution $c_1 = 4/3$ and $c_2 = -1/3$, so the solution of the initial value problem is

$$y(x) = (4/3)e^x - (1/3)e^{-2x}.$$

(ii) The characteristic equation is

$$\lambda^2 + 2\lambda + 4 = 0,$$

with $A^2 - 4B = -12$, so this is Case (II) with $\alpha = -1$ and $\beta = \sqrt{3}$. The general solution is

$$y(x) = e^{-x}[c_1\cos(x\sqrt{3}) + c_2\sin(x\sqrt{3})].$$

The initial condition $y(0) = 2$ is satisfied if $2 = c_1$, while the initial condition $y'(0) = 1$ is satisfied if

$$1 = -2 + c_2\sqrt{3}.$$

Solving these equations gives $c_1 = 2$ and $c_2 = \sqrt{3}$, so the solution of the initial value problem is

$$y(x) = e^{-x}[\sqrt{3}\sin(x\sqrt{3}) + 2\cos(x\sqrt{3})].$$

(iii) The characteristic equation is

$$\lambda^2 + 4\lambda + 4 = 0,$$

with $A^2 - 4B = 0$, so this is Case (III) with $\mu = -2$. The general solution is

$$y(x) = (c_1 + c_2x)e^{-2x}.$$

Using the initial condition $y(0) = 3$ shows that $3 = c_1$, whereas the initial condition $y'(0) = 1$ will be satisfied if

$$1 = -6 + c_2.$$

Solving these equations gives $c_1 = 3$ and $c_2 = 7$, so the solution of the initial value problem is

$$y(x) = (3 + 7x)e^{-2x}.$$

■

We now formulate the fundamental existence and uniqueness theorem for the homogeneous linear second order constant coefficient differential equation (1). This is a special case of a more general theorem that will be quoted later.

THEOREM 6.1

existence and uniqueness of solutions

Existence and uniqueness of solutions of homogeneous second order constant coefficient equations Let differential equation (1) have two linearly independent solutions $y_1(x)$ and $y_2(x)$. Then, for any $x = x_0$ and numbers μ_1 and μ_2, a unique solution of (1) exists satisfying the initial conditions

$$y(x_0) = \mu_0, \quad y^{(1)}(x_0) = \mu_1.$$

Proof The existence of the solutions $y_1(x)$ and $y_2(x)$ was established when the cases (I), (II), and (III) were examined. The nonvanishing of the determinant Δ in (9) showed c_1 and c_2 to be uniquely determined by the given initial conditions when the roots are real and distinct, so the solution of the initial value problem is also unique. An examination of the form of the determinant Δ in cases (II) and (III) establishes the uniqueness of the solution in the remaining two cases, though the details are left as an exercise. ■

two-point boundary conditions

A different type of problem that can arise with second order equations occurs when the solution is required to satisfy a condition at two distinct points $x = a$ and $x = b$, instead of satisfying two initial conditions. Problems of this type are called **two-point boundary value problems**, because the points a and b can be regarded as boundaries between which the solution is required, and at which it must satisfy given **boundary conditions**. Problems of this type occur in the study of the bending of beams that are supported in different ways at each end, and elsewhere (see Section 8.10).

Typical two-point boundary value problems involve either the specification of $y(x)$ at $x = a$ and at $x = b$, or the specification of $y(x)$ at one boundary and $y'(x)$ at the other one. The most general two point boundary value problem involves finding a solution in the interval $a < x < b$ such that

$$y'' + Ay' + By = 0,$$

subject to the boundary condition at $x = a$

$$\alpha y(a) + \beta y'(a) = \mu,$$

and the boundary condition at $x = b$

$$\gamma y(b) + \delta y'(b) = K,$$

where α, β, γ, δ, μ, and K are known constants.

EXAMPLE 6.5

Solve the two-point boundary value problem

$$y'' + 2y' + 17y = 0, \quad \text{with } y(0) = 1 \text{ and } y'(\pi/4) = 0.$$

Solution The characteristic equation is

$$\lambda^2 + 2\lambda + 17 = 0$$

with the complex roots $\lambda_1 = -1 + 4i$ and $\lambda_2 = -1 - 4i$, so the general solution is

$$y(x) = e^{-x}[c_1 \cos 4x + c_2 \sin 4x].$$

At the boundary $x = 0$ the general solution reduces to $1 = c_1$, whereas at the boundary $x = \pi/4$ it reduces to $0 = -e^{-\pi/4} + 4c_2 e^{-\pi/4}$, showing that $c_2 = 1/4$. So the solution of the two-point boundary value problem is

$$y(x) = e^{-x}\left[\cos 4x + \frac{1}{4}\sin 4x\right], \quad \text{for } 0 < x < \pi/4.$$ ■

Summary

This section introduced the homogeneous linear second order constant coefficient equation and explained the importance of the linear independence of solutions. It showed how

for this second order equation the general solution can be expressed as a linear combination of the two linearly independent solutions that can always be found. The form of the two linearly independent solutions was shown to depend on the relationship between the roots of the characteristic equation. A fundamental existence and uniqueness theorem was given and the nature of a simple two-point boundary value problem was explained.

EXERCISES 6.1

In Exercises 1 through 4 test the given pairs of functions for linear independence or dependence over the stated intervals.

1. **(a)** $\sinh^2 x, \cosh^2 x$, for all x.
(b) $x + \ln|x|, x + 2\ln|x|$, for $|x| > 0$.
(c) $1 + x, x + x^2$, for all x.

2. **(a)** $\sin x, \cos x$, for all x.
(b) $\sin x \cos x, \sin 2x$, for all x.
(c) e^{2x}, xe^{2x}, for all x.

3. **(a)** $|x|x^2, x^3$, for $-1 < x < 1$.
(b) $\sin x, \tan x$, for $-\pi/4 \le x \le \pi/4$.
(c) $x|x|, x^2$, for $x \ge 0$.

4. **(a)** $\sin x, |\sin x|$, for $\pi \le x \le 2\pi$.
(b) $x^3 - 2x + 4, -4x^3 + 8x - 16$, for all x.
(c) $x + 2|x|, x - 2|x|$ for all x.

Find the general solution of the differential equations in Exercises 5 through 20.

5. $y'' + 3y' - 4y = 0$.
6. $y'' + 2y' + y = 0$.
7. $y'' - 2y' + 2y = 0$.
8. $y'' + 2y' + 2y = 0$.
9. $y'' + 2y' - 3y = 0$.
10. $y'' + 5y' + 4y = 0$.
11. $y'' + 6y' + 9y = 0$.
12. $y'' - 2y' + 4y = 0$.
13. $y'' - 4y' + 5y = 0$.
14. $y'' + 3y' + 3y = 0$.
15. $y'' + 6y' + 25y = 0$.
16. $y'' - 4y' + 20y = 0$.
17. $y'' + 5y' + 4y = 0$.
18. $y'' + 4y' + 5y = 0$.
19. $y'' - 3y' + 3y = 0$.
20. $y'' + y' + y = 0$.

Solve initial value problems in Exercises 21 through 28 using the method of this section, and confirm the solutions for even numbered problems by using computer algebra.

21. $y'' + 5y' + 6y = 0$, with $y(0) = 1, y'(0) = 2$.
22. $y'' + 4y' + 5y = 0$, with $y(0) = 1, y'(0) = 3$.
23. $y'' + 2y' + 2y = 0$, with $y(0) = 3, y'(0) = 1$.
24. $y'' + 6y' + 8y = 0$, with $y(0) = 1, y'(0) = 0$.
25. $y'' - 5y' + 6y = 0$, with $y(0) = 2, y'(0) = 1$.
26. $y'' - 3y' + 3y = 0$, with $y(0) = 0, y'(0) = 2$.
27. $y'' - 3y' - 4y = 0$, with $y(0) = -1, y'(0) = 2$.
28. $y'' - 2y' + 3y = 0$, with $y(0) = 1, y'(0) = 0$.

Solve the boundary value problems in Exercises 29 through 36 using the method of this section, and confirm the solutions for even-numbered problems by using computer algebra.

29. $y'' + 4y' + 3y = 0$, with $y(0) = 1, y'(1) = 0$.
30. $y'' + 4y' + 4y = 0$, with $y(0) = 2, y'(1) = 0$.
31. $y'' + 6y' + 9y = 0$, with $y(-1) = 1, y'(1) = 0$.
32. $y'' + 4y' + 5y = 0$, with $y(-\pi/2) = 1, y'(\pi/2) = 0$.
33. $y'' + 2y' + 26y = 0$, with $y(0) = 1, y'(\pi/4) = 0$.
34. $y'' + 2y' + 26y = 0$, with $y(0) = 0, y'(\pi/4) = 2$.
35. $y'' + 5y' + 6y = 0$, with $y(0) = 0, y'(1) = 1$.
36. $y'' + 2y' - 3y = 0$, with $y(0) = 1, y'(1) = 1$.

Theorem 6.1 ensures the existence and uniqueness of solutions of initial value problems for the differential equation in (1), but does not apply to two-point boundary value problems that may have no solution, a unique solution or infinitely many solutions. In Exercises 37 and 38 use the general solution of

$$y'' + y = 0$$

to find if a solution exists and is unique, exists but is nonunique, or does not exist for each set of boundary conditions.

37. **(a)** $y(0) = 0, y(\pi) = 0$.
(b) $y(0) = 1, y(2\pi) = 2$.
(c) $y'(0) = 1, y(\pi/4) = \sqrt{2}$.

38. **(a)** $y(0) = 1, y(\pi/2) = 1$.
(b) $y(0) = 0, y'(\pi) = 0$.
(c) $y'(0) = 0, y'(\pi) = 0$.

39. For what values of λ will the following two-point boundary value problem have infinitely many solutions, and what is the form of these solutions:

$$y'' + \lambda^2 y = 0, \quad \text{with } y(0) = 0, y(\pi) = 0.$$

40. A particle moves in a straight line in such a way that its distance x from the origin at time t obeys the differential equation $x'' + x' + x = 0$. Assuming it starts from the origin with speed 30 ft/sec, what will be its distance from the origin, its speed, and its acceleration after $\pi/\sqrt{3}$ seconds?

41. The angular displacement θ of a damped simple pendulum obeys the equation $\theta'' + 2\mu\theta' + (\mu^2 + p^2)\theta = 0$,